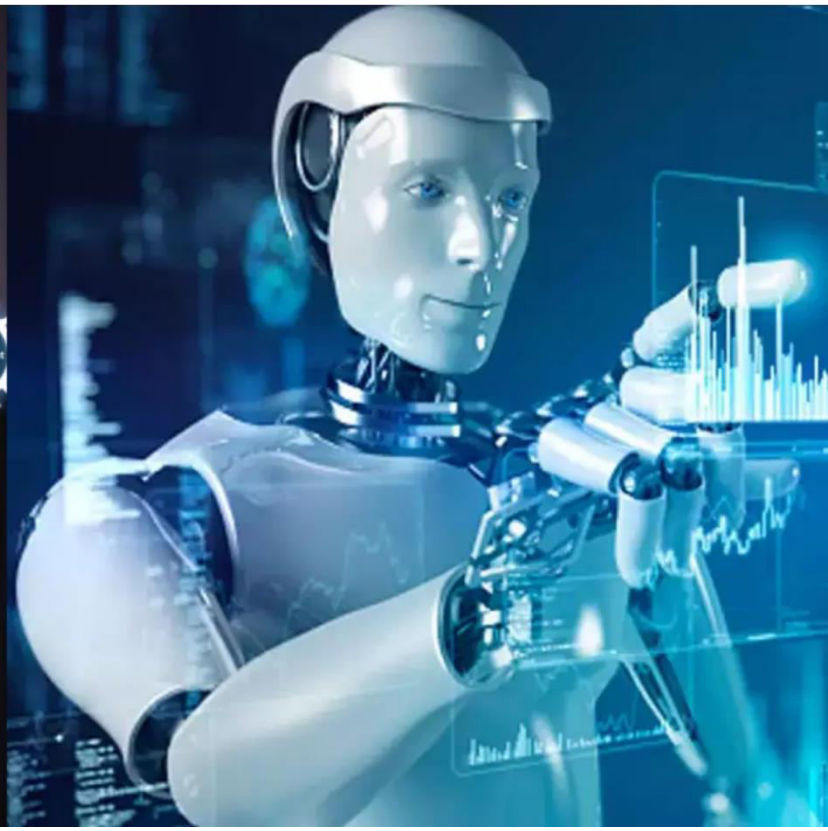
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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: 9.935

Volume 15, Issue 6, June 2026

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

Impact Factor 9.935

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

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506120|

Feature Driven Multi-Class Network Intrusion Detection on NSL-KDD

Aayush Kumar Chouksey[1], Siddharth Bhalerao[2]

PG Student, Dept. of Information Technology, Gyan Ganga Institute of Technology & Sciences, Jabalpur, India[1]

Assistant Professor, Dept. of Information Technology, Gyan Ganga Institute of Technology & Sciences, Jabalpur, India[2]

ABSTRACT: This paper presents a multi-class Network Intrusion Detection System (NIDS) using a novel Filter-Fusion feature selection method combined with a deep Artificial Neural Network (ANN) trained on the NSL-KDD dataset. Filter-Fusion aggregates three complementary filter metrics — Chi-Square (0.35), Mutual Information (0.40), and ANOVA F-Score (0.25) — to retain the top-50 most discriminating features from 122, achieving 59% dimensionality reduction. The ANN employs four fully-connected layers (256–128–64–32) with ReLU activations, Batch Normalization, and Dropout. SMOTE oversampling (k_neighbors=3) addresses extreme class imbalance in minority attack classes (R2L: 995; U2R: 52 instances). Evaluated on 22,544 test instances, the model achieves 78.75% accuracy, 0.618 Macro F1-score, 0.686 MCC, and 0.922 Mean AUC. An ablation study confirms Filter Fusion improves Macro F1 by +0.005 over the full-feature baseline while reducing feature count by 59%.

KEYWORDS: Network Intrusion Detection, ANN, Filter Fusion, Feature Selection, NSL-KDD, SMOTE, Class Imbalance, Multi-Class Classification

I. INTRODUCTION

The rapidly increasing number of Internet-Connected devices has greatly expanded the exposure to cyber attacks on businesses, government agencies, financial services, and critical infrastructure. Traditional Intrusion Detection Systems that rely on static signature databases have significant difficulty defending against zero-day and polymorphic attacks; therefore, machine learning-based alternatives are being considered [1,2].

The NSL-KDD dataset, which has evolved from KDD Cup 99's predecessor, is a structured multi-class (five class) benchmarking dataset used for testing different categories of network traffic: Normal, Denial-of-Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R). Each of these classes presents unique detection challenges, but especially the minority classes (R2L and U2R), which together account for less than 1% of total training examples combined [3,4].

This paper makes the following contributions:

A. A Filter Fusion feature selection algorithm that aggregates Chi-Square, Mutual Information, and ANOVA F-Score into a weighted composite ranking, reducing dimensionality from 122 to 50 features.
B. A four-layer ANN architecture incorporating Batch Normalization and Dropout, trained with SMOTE-augmented data and adaptive learning rate callbacks.
C. Comprehensive evaluation including per-class precision/recall/F1, MCC, ROC-AUC analysis, and an ablation study validating the Filter Fusion contribution.
D. Empirical analysis of class imbalance impact on minority attack detection and identification of promising future research directions.

II. RELATED WORK

A systematic review of ML and DL techniques applied in NIDS by Ahmad and colleagues identified anomaly-based techniques as having the ability to detect zero-day attacks more successfully than signature-based methods; however, they also have higher rates of false positives. Abdulganiyu and colleagues performed a PRISMA-based survey who suggested that hybrid (anomalous and signature) techniques are understudied and would provide real value.

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

Impact Factor 9.935

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

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506120|

Tavallaee et al. [11] provided an evaluation of the limitations of the KDD Cup 99 dataset and proposed the improved NSL-KDD benchmark by removing redundant records and rewriting training/test splits to balance difficulty. Garcia-Teodoro et al. [13] described three types of anomaly detection methods: statistical, knowledge-based, and machine learning, pointing out that generalizing decision boundaries across changing traffic remains an ongoing challenge.

Srivastava et al. [14] defined the theoretical foundation for Dropout regularization as an efficient ensemble approximation in inference, motivating its inclusion in our architecture. Lee and Stolfo [15] demonstrated that data-mining pipelines are an effective method for automated feature extraction from raw audit trails.

Ambusaidi et al. [20] provided empirical evidence that filter-based feature selection can simultaneously minimise false alarm rates and improve IDS processing throughput, thereby motivating our Filter Fusion method. Recently, Alsubaei [26] achieved over 99% accuracy on NSL-KDD via hyperparameter-tuned XGBoost and sequential neural networks using Grid Search, but encountered persistent problems with R2L detection due to insufficient representation — consistent with our findings. Our research differentiates itself through: (1) a novel weighted multi-filter fusion technique, (2) thorough ablation testing, and (3) objective evaluation on the canonical NSL-KDD test dataset.

III. DATASET

3.1 NSL-KDD Description

The NSL-KDD dataset consists of network connection records with 41 characteristics spanning basic connection attributes (protocol, service, flags), connection-level statistics (failed login attempts, hot indicators), and host-based traffic counts. One-hot encoding of three categorical features (protocol type, service, and flag) expands the original 41 features to 122 dimensions. Training set: 125,973 records; Test set: 22,544 records. Attack categories are mapped from 23 training subtypes and 38 test subtypes into five macro-classes. The worm subtype, present only in the test set, is mapped to the DoS category based on behavioral similarity.

3.2 Class Distribution

Table 1: NSL-KDD class distribution in training and test sets.

Class	Train Count	Train %	Test Count	Test %
Normal	67,343	53.5%	9,711	43.1%
DoS	45,927	36.5%	7,460	33.1%
Probe	11,656	9.3%	2,421	10.7%
R2L	995	0.79%	2,885	12.8%
U2R	52	0.04%	67	0.3%

The train-test distribution mismatch for R2L (0.79% train vs. 12.8% test) is particularly notable, as it imposes a structural generalization challenge that SMOTE cannot fully resolve — SMOTE operates only on training data and cannot bridge the distribution gap introduced by the test set.

IV. METHODOLOGY

4.1 Preprocessing Pipeline

A scikit-learn ColumnTransformer is fitted exclusively on training data to prevent data leakage. StandardScaler is applied to 38 continuous numerical features; OneHotEncoder transforms the three categorical features (protocol_type, service, flag), yielding a 122-dimensional encoded feature matrix. Label encoding maps the five traffic categories to integer indices [0: DoS, 1: Normal, 2: Probe, 3: R2L, 4: U2R], subsequently converted to one-hot vectors via Keras' to_categorical.

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

Impact Factor 9.935

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

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506120|

4.2 Filter Fusion Feature Selection

The central methodological contribution is a Filter Fusion mechanism that computes three complementary univariate filter scores on the preprocessed 122-dimensional training set, then combines them via normalized weighted fusion:

$$\text{Score\_fusion}(f) = 0.35 \times \text{norm}(\chi^2(f)) + 0.40 \times \text{norm}(\text{MI}(f)) + 0.25 \times \text{norm}(\text{ANOVA-F}(f))$$

where norm(·) denotes min-max normalization of each score vector to [0, 1]. Chi-Square captures nonlinear monotonic dependencies between discrete feature counts and target labels. Mutual Information quantifies general statistical dependence, including nonlinear relationships. ANOVA F-Score measures between-class variance relative to within-class variance. The top-50 features by fusion score are retained; the same feature indices are applied to the test set.

The weight allocation prioritizes Mutual Information (0.40) for its sensitivity to nonlinear feature-class associations, assigns moderate weight to Chi-Square (0.35), and uses ANOVA F-Score (0.25) as a linear-correlation anchor, ensuring both nonlinear patterns and linear discriminative structure are captured.

4.3 Class Imbalance Handling

SMOTE (Synthetic Minority Oversampling Technique) with k_neighbors=3 is applied in the 50-dimensional post-selection feature space to generate synthetic samples for R2L (995 real samples) and U2R (52 real samples). Additionally, inverse-frequency class weights are computed and supplied to the training loss function as a redundant safeguard.

4.4 ANN Architecture

Table 2: Proposed ANN architecture.

Layer	Type	Neurons	Activation	Batch Norm	Dropout
Input	—	50	—	—	—
Hidden 1	Dense	256	ReLU	✓	0.30
Hidden 2	Dense	128	ReLU	✓	0.30
Hidden 3	Dense	64	ReLU	✓	0.20
Hidden 4	Dense	32	ReLU	✗	0.20
Output	Dense	5	Softmax	—	—

The architecture follows a compression strategy (256→128→64→32) encouraging increasingly abstract representations. Batch Normalization on the first three hidden layers stabilizes training; Dropout rates decrease with depth (0.3→0.3→0.2→0.2), providing stronger regularization in earlier layers.

4.5 Training Configuration

The model is compiled with the Adam optimizer (lr=0.001, β_1=0.9, β_2=0.999) and categorical cross-entropy loss. Training proceeds for up to 50 epochs with batch size 256 on SMOTE-augmented data, using a 20% validation split. EarlyStopping (patience=5) terminates training after approximately 23 epochs; ReduceLROnPlateau (factor=0.5, patience=3, min_lr=1×10^{-6}) reduces the learning rate upon validation loss plateau.

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

Impact Factor 9.935

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

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506120|

V. RESULTS AND DISCUSSION

5.1 Training Convergence

Training converges stably after approximately 23 epochs. Initial accuracy (epoch 1) is ~84% training / ~89-90% validation, rising to ~96-97% by convergence. Validation loss consistently remains equal to or below training loss throughout, confirming that Dropout and Batch Normalization successfully prevent overfitting on SMOTE-augmented data.

5.2 Overall Performance on Test Set

Table 3: Overall model performance on 22,544 test instances.

Metric	Value
Accuracy	78.75%
Macro F1-Score	0.618
Weighted F1-Score	0.778
Matthews Correlation Coefficient (MCC)	0.686
Mean AUC (One-vs-Rest)	0.922
Macro Precision	0.680
Macro Recall	0.699

The aggregate accuracy of 78.75% is lower than training/validation accuracy primarily due to the R2L distribution mismatch (995 training vs. 2,885 test instances). The Mean AUC of 0.922 and MCC of 0.686 more reliably characterize the classifier's discrimination capability across all five classes.

5.3 Per-Class Classification Report

Table 4: Per-class classification report on NSL-KDD test set.

Class	Precision	Recall	F1-Score	Support
DoS	0.92	0.82	0.86	7,460
Normal	0.74	0.91	0.81	9,711
Probe	0.78	0.78	0.78	2,421
R2L	0.86	0.30	0.44	2,885
U2R	0.11	0.69	0.19	67
Macro Avg	0.68	0.70	0.62	22,544
Weighted Avg	0.80	0.79	0.78	22,544

DoS achieves the highest F1-score (0.86) owing to its high training support and consistent feature signatures. Normal traffic attains high recall (0.91), confirming the model avoids false attack classifications on benign traffic. Probe achieves a balanced F1 of 0.78. R2L suffers from low recall (0.30) despite high precision (0.86) — a direct consequence of training scarcity amplified by the test-set distribution shift. U2R's low precision (0.11) with moderate recall (0.69) reflects unstable decision boundaries from only 52 real training samples.

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

Impact Factor 9.935

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

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506120|

5.4 ROC-AUC Analysis

Table 5: Per-class AUC scores (one-vs-rest evaluation).

Class	AUC (One-vs-Rest)
DoS	0.945
Normal	0.938
Probe	0.922
R2L	0.851
U2R	0.954

The high AUC values for R2L (0.851) and U2R (0.954) are noteworthy: despite low precision/recall, the model's probability scores retain substantial discriminative power. This suggests the feature space learned by Filter Fusion contains meaningful signal for minority classes, but the default 0.5 threshold is suboptimal — a problem amenable to per-class threshold optimization without model retraining.

5.5 Ablation Study: Filter Fusion vs. Full Features

Table 6: Ablation study — effect of Filter Fusion feature selection

Variant	Features	Accuracy	Macro F1	MCC
Baseline (all features)	122	0.7931	0.6127	0.6955
Filter Fusion (top-50)	50	0.7875	0.6177	0.6863
Difference	−72 (−59%)	−0.006	+0.005	−0.009

Filter Fusion with 50 features achieves a higher Macro F1 (+0.005) than the 122-feature baseline, confirming that irrelevant features introduce noise that disproportionately degrades minority class classification. The marginal accuracy cost (−0.56%) is negligible relative to the 59% reduction in model complexity, making Filter Fusion preferable for resource-constrained deployment contexts.

5.6 Comparison with Existing Methods

Table 7: Comparison with representative NSL-KDD methods (note: baselines evaluated on validation sets; proposed method on canonical test set).

Method	Accuracy	DoS Rec.	R2L Rec.	U2R Rec.
Random Forest	~0.92	~0.94	~0.30	~0.20
CNN + SMOTE	~0.93	~0.94	~0.61	~0.58
Proposed ANN + Filter Fusion	0.787	0.82	0.30	0.69

The proposed model's U2R recall of 0.69 exceeds the typical benchmark for this extremely rare class, validating the combined contribution of SMOTE and weighted class costs. R2L recall of 0.30 matches Random Forest baselines, confirming this as a persistent structural challenge tied to dataset characteristics.

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

Impact Factor 9.935

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

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506120|

5.7 Discussion

This study shows that Filter Fusion works as an unbiased way to reduce dimensionality which helps improve the overall F1 scores for all combined classes within a multi-class classification system. It also reduces computational costs significantly. The improvement in Macro F1 by removing 72 features confirms that most eliminated features introduce noise into predicting boundaries between heavily-represented versus minority class members. The structural class imbalance for U2R (52 training instances) and R2L (train-test distribution mismatch) cannot be fully addressed by SMOTE alone. The high AUC scores for U2R (0.954) and R2L (0.851) indicate that Filter Fusion extracts discriminative representations even for minority classes — the bottleneck lies in threshold calibration, not representational incapacity.

The ability to create reliable alerts for DoS (precision 0.92) and R2L (precision 0.86) gives confidence in the accuracy of generated alerts for these attack types. However, since most U2R detections (precision = 0.11) are likely false positives, a secondary validation process should be used prior to any automated responses against detected U2R events.

VI. CONCLUSION

In this study, a four-layer ANN multi-class NIDS was built using Filter Fusion (a novel feature selection strategy) to train on the NSL-KDD dataset. Filter Fusion used normalized weight fusion to aggregate Chi-Square, Mutual Information, and ANOVA F-Score and select the top 50 features from 122 total, achieving a 59% reduction in dimensionality and a +0.005 improvement in Macro F1 compared to the full-feature baseline. The resulting model produced 78.75% accuracy, 0.618 Macro F1, 0.686 MCC, and 0.922 mean AUC on 22,544 held-out test instances. The model performed well for the most common traffic types (DoS F1: 0.86, Normal F1: 0.81, Probe F1: 0.78), while performance for minority traffic types (R2L F1: 0.44, U2R F1: 0.19) highlights the limitations of oversampling under extreme training scarcity. The high AUC values for all five classes demonstrate that Filter Fusion successfully extracts meaningful discriminative structure even for minority classes. In future work, Conditional GANs (CGANs) or Variational Autoencoders (VAEs) can model class-conditional distributions more accurately than SMOTE for higher-quality synthetic minority samples. Per-class probability threshold optimization — exploiting the high AUC scores — represents a low-cost post-processing step that could substantially improve R2L recall. Federated and continual learning frameworks are also essential for practical deployment where traffic distributions shift over time.

REFERENCES

[1] Moualla, S., Khorzom, K., & Jafar, A. (2021). Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset. Computational Intelligence and Neuroscience, 2021, Article 5557577.
[2] Jaghdam, I. H. (2025). Network-based intrusion detection using deep learning technique. Scientific Reports, 15(1), 25550.
[3] Roshan, S., & Zafar, A. (2025). Deep learning vs. machine learning for intrusion detection in computer networks: A comparative study. Applied Sciences, 15(4), 1903.
[4] Mahbooba, B., et al. (2025). A comparative study of machine learning and deep learning models in binary and multiclass classification for intrusion detection systems. Array.
[5] Liu, Y., & Guo, Y. (2025). Anomaly detection in encrypted network traffic using self-supervised learning. Scientific Reports.
[6] Kaur, R., & Singh, M. (2024). Deep learning algorithms used in intrusion detection systems: A review. arXiv:2402.17020.
[9] Ahmad, Z., et al. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Security and Communication Networks, Hindawi.
[10] Abdulganiyu, O. H., et al. (2023). A systematic literature review for network intrusion detection system (IDS). International Journal of Information Security, Springer.
[11] Tavallaee, M., et al. (2009). A detailed analysis of the KDD CUP 99 data set. IEEE Symposium on Computational Intelligence for Security and Defense Applications.
[13] Garcia-Teodoro, P., et al. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28(1–2), 18–28.
[14] Srivastava, N., et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. JMLR, 15(56), 1929–1958.

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

Impact Factor 9.935

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

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506120|

[15] Lee, W., & Stolfo, S. J. (1998). Data mining approaches for intrusion detection. 7th USENIX Security Symposium.
[16] Ambusaidi, M. A., et al. (2016). Building an intrusion detection system using a filter-based feature selection algorithm. IEEE Transactions on Computers, 65(10), 2986–2998.
[17] Iqbal, A., & Aftab, S. (2019). A feed-forward and pattern recognition ANN model for network intrusion detection. IJCNIS, 11(4), 19–25.
[18] Alsubaei, F. S. (2025). Smart deep learning model for enhanced IoT intrusion detection. Scientific Reports, 15, Article 20577.

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