Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

Academia Open

By Universitas Muhammadiyah Sidoarjo

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

Table Of Contents

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

EDITORIAL TEAM

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

Article information

Check this article update (crossmark)

Check this article impact (*)

Save this article to Mendeley

(*) Time for indexing process is various, depends on indexing database platform

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

Network Traffic Classification Using Machine Learning Techniques: An Applied Flow-Based Study

Nahla Flayyih Hasani, nahla.flayyih@gmail.com (*)

College of Basic Education, University of Sumer-Rifai, Thi-Qar, Iraq

(*) Corresponding author

Abstract

General Background Network data traffic rating represents a core task in modern network engineering for service quality control and bandwidth management. **Specific Background** Traditional approaches relying on deep packet inspection and port-based categorization face severe limitations due to the proliferation of dynamic ports and widespread encryption protocols such as TLS and HTTPS. **Knowledge Gap** Although automated algorithms offer alternative solutions, comprehensive applications detailing complete processing pipelines that systematically compare multiple predictive models without inspecting actual packet contents remain scarce. **Aims** This study designs an integrated flow-based data processing pipeline to extract nine statistical properties for categorizing six data movement classes using four distinct automated algorithms. **Results** Utilizing a simulated dataset of 3000 flows, experimental outcomes revealed that the aggregate tree model attained the highest precision of 97.87 percent and a macro F1-score of 97.87 percent, significantly outperforming Decision Tree, Support Vector Machine, and KNN models. **Novelty** This research demonstrates the explicit superiority of ensemble techniques in data categorization by identifying byte rate, mean packet length, and total bytes as the most decisive flow-level metrics, entirely independent of payload decryption. **Implications** Relying on extracted statistical properties presents a highly reliable methodology for managing modern encrypted infrastructures, suggesting that administrators should prioritize aggregate machine learning systems to optimize resource allocation and detect unwanted activities.

Highlights:

- Aggregate tree models attained a 97.87 percent precision rate across six data movement categories.
- Payload decryption is unnecessary when utilizing statistical properties such as byte rates and packet lengths.
- Machine learning pipelines successfully categorized video streaming, web browsing, and voice communications without port dependency.

Keywords: Machine Learning, Data Movement Categorization, Statistical Properties, Deep Packet Inspection, Decision Tree Models

Published date: 2026-08-18

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

Introduction

Network data traffic rating is one of the most important and complex tasks in the field of modern network engineering, with several basic functions such as bandwidth management, service quality control (QOS), resource allocation between different applications, as well as detecting malicious or unwanted activities within network. Historic systems have relied on two main methods: Port-Based Classification and Deep Packet Inspection - DPI. However, both methods faced fundamental restrictions with the development of the Internet; The propagation of dynamic ports and applications that hide behind non-standard ports has weakened the reliability of port-based classification [1]. While the package checking content has become practically unavailable with the widespread spread of encryption protocols such as TLS and HTTPS, which are now covering the vast majority of global Internet traffic [2] [3].

With these challenges, machine learning has emerged as a promising alternative based on the extraction of statistical properties from data movement (such as package lengths, intervals between their arrival, flow duration, and the number of bytes transferred) without the need for Read the actual package content or exclusive dependence on the port number. Recent studies have proven that this approach is capable of achieving high classification accuracy even in the presence of fully encrypted data traffic [4][5][6], which has made it an active and continuous subject of research in recent years.

This research aims to provide an integrated application study that is not limited to the literature review, but includes an actual design and implementation of a complete data processing and classification pipeline, from the definition of properties, through pretreatment, and ending with the training of four. Automated learning models that are commonly used in this field, and are evaluated in comparison with multiple performance standards.

Specifically, this research seeks to answer the following research questions:

- How effective are the flow-based features in distinguishing different data traffic?
    - Which of the four studied classification algorithms (Decision Tree, Random Forest, SVM, KNN) achieve the best balance between accuracy and stability?
    - What are the most influential characteristics of the classification decision, and how can this be explained by the network?

The structure of this research is divided into six main sections: the second section reviews the relevant work, and the third section presents the proposed methodology, including the description of data, characteristics and classification algorithms. It will discuss it, and the sixth section concludes with the most important future conclusions and recommendations.

2. Related Work

AZAB and others [1] presented a comprehensive survey that compiled networks movement classification techniques and data sets, and discussed in-depth the challenges related to the diversity of protocols and cargo encryption. On the other hand, Lopez-Martin et al. [7] reviewed the deep learning applications in network traffic control and analysis (NTMA), with a special focus on the comparison between precision and computational complexity in actual production environments.

In more recent work, Feng et al. [3] provided a detailed survey of fine-grained internet traffic analysis techniques, discussing the limitations of traditional statistical methods in the face of the rapidly growing diversity of modern applications.

In the context of combining machine learning with Software-Defined Networking (SDN), Serag et al. [8] and Salau and Beyene [9] presented two applied studies which employed algorithms like AdaBoost, random forests, decision trees, logistic regression and SVM for classify protocols such as Telnet, DNS and Voice within an SDN environment; meanwhile, Sobh et al. [10] proposed a similar model aimed at improving Quality of Service (QoS) via scenarios of binary and multi-class classification, compared to linear, non-linear, and hybrid models.

Regarding deep learning contexts for encrypted traffic classification, Shi et al. [5] proposed the BFCN model, which integrates representations of BERT with Convolutional Neural Networks (CNNs). Meanwhile, Elshewey and Osman [6] produced a stacked deep ensemble framework combining GRUs, CNNs, RNNs, DNNs and LSTMs to classify encrypted HTTPS traffic, achieving an accuracy exceeding 99%. Additionally, Eslami and Hamouda [11] produced a hybrid framework leveraging self-supervised learning and confident learning for reduce reliance on manually classified data; this framework was evaluated on well-known datasets, including ISCX VPN-nonVPN.

In a recent comparative study, Antari et al. [4] employed classical models

(Random Forest, XGBoost, Decision Tree) alongside a Transformer model and large language models (Gemini, GPT-4o); the results produced the superiority of the Transformer model, achieving an accuracy of 98.95%. Meanwhile, Saqib et al. [12] proposed an in-network adaptive classifier aimed at reducing misclassification rates to enhance Quality of Service (QoS). Finally, Pekár et al. [13] proposed a comprehensive tutorial on building systems of flow-based classification, feature engineering, steps of covering flow measurement and leakage-resistant experimental design which largely informed the methodology of the current study.

In a context similar to the current application, Alshammari and Aleroud [14] compared between Random Forest, Decision Tree, KNN and SVM algorithms to classifying network traffic of smart city; they concluded that the Decision Tree algorithm outperformed the others with an accuracy of 99.18%, providing a direct point of comparison with the results of this research.

The current search for these works is distinguished by its focus on building a complete and reproductable application tube, with a systematic comparison of four algorithms within a standardized experimental environment and multiple evaluation criteria (accuracy, weighted accuracy, and recall. F1 score, cross-check, rather than just displaying individual results or purely literary review.

3. Proposed Methodology

This section describes the applied methodology of the research, which consists of four sequential stages: (a) the description of the flow data set, (b) the extraction of statistical properties, (c) the pre-processing of the data, (d) the selection of classification algorithms and criteria evaluation.

3.1 Dataset Description

Due to the limitations of access to the Internet within the implementation environment of this search, it is not possible to download popular reference datasets such as ISCX VPN-Nonvpn [11] or CICIDS2017 directly.

To avoid a theoretical presentation, a simulated flow dataset was built statistically calibrate to mimic the properties of actual data movement documented in reference literature [1][14][7] For each of the six categories of applications; The average and standard deviation of each property (such as beam length, flow duration, and beam interval) were determined based on the values reported in those studies of each application type, with NOISE and deliberate interference between Items to avoid getting a simplified and unrealistic rating task. Table 1 shows the distribution of the six items used.

Traffic Class	description	Number of flows
Web	Standard web browsing (HTTP/HTTPS)	500
Video_Streaming	Video streaming (such as YouTube/Netflix)	500
VoIP	Voice calls over the Internet	500
File_Transfer	Transferring large files (FTP/download)	500
Email	Email correspondence (SMTP/IMAP)	500
Gaming	Online interactive games	500
Total	—	**3000**

Table 1.
Table 1 illustrates the distribution of the six categories used.

It is important to note that relying on statistically calibrated data aims to provide a fully controlled experimental environment one that is entirely reproducible for testing the integrity of the processing pipeline and the proposed classification; this is clearly identified as a study limitation in Section 5, alongside an explicit recommendation to re-verify the results using real-world data in future work.

3.2 Extraction of Statistical Flow Features

Nine statistical features were extracted for each network flow, defined as the set of packets exchanged between specific source and destination addresses within a single time window. These features are summarized in Table 2; they were selected depended on the most widely used and effective features found in the literature on flow classification [4][15][14], as they do not require inspection of the packet payload and thus remain valid even with fully encrypted data traffic.

Property name	**description**
mean_packet_length	Average packet length (bytes) within the flow
flow_duration_s	Total flow duration (seconds)
mean_iat_s	Average time interval between the arrival of consecutive packets
fwd_packets	Number of packets in the forward direction(Client → Server)
bwd_packets	Number of packets in the reverse direction (Server → Client)
total_bytes	Total number of bytes transferred in the stream
is_udp	A binary property indicating the transport protocol (UDP = 1 TCP = 0)
packet_rate	Rate of packets sent per second
byte_rate	Rate of bytes transferred per second

Table 2.
Table 2: Statistical characteristics extracted from each network flow

3.3 Data Preprocessing

The preprocessing stage comprised the following steps: First, label encoding was performed to convert the names of the six categories into numerical values. Second, the data was split into training and testing sets, using a 75% to 25% ratio, while maintaining a stratified split across the categories to ensure balanced representation. Third, feature scaling was applied using the Standard Scalar; this normalization was performed exclusively for models sensitive to data scale, namely SVM and KNN, whereas raw, raw features were used for the Decision Tree and Random Forest models, as these are insensitive to value scaling.

3.4 Classification Algorithms Used

Four classification algorithms representing different currents were selected under Supervised Learning, and are most employed in network movement rating literature [8][14][9][10]:

- Decision Tree: An explanatory model based on the repeated division of data according to Thresholds for each property, with a maximum depth specified by 10 levels to avoid overfitting.
- Random Forest: An aggregate model that combines 200 independent decision trees based on random subsamples of data and properties, with a maximum depth of 12 levels, with the aim of improving stability and reducing contrast.
- Supporting vector machine (SVM): using a radial kernel (RBF) with the coefficient of C = 10, an effective algorithm in medium-dimensional spaces when there are non-linear separation boundaries between the cultivars.
- KNN: Using seven neighbors (K=7), a lazy learning based on the majority vote among the closest data points in the unified properties space.

3.5 Evaluation Metrics

To comprehensively evaluate the performance of the four models, four standard criteria were adopted: accuracy, macro-averaged precision, macro-averaged recall, and total F1 score. (Macro F1-Score); this is to deal fairly with all six items, without weighting the most representative items. In addition, K-Fold stratification (5 folds) was used on training data only, with the aim of verifying the stability of each model's performance and reducing the probability of bias resulting from a single randomized split of data.

4. Experimental Setup and Practical Implementation

The full application aspect of this research is carried out using the Python programming language (Version 3), based on the Scikit-Learn to build and evaluate models, and the Pandas and Numpy for Data Processing and Statistical Generation. The MatPlotLib library to produce illustrations. The implementation tube consisted of the following stages in sequence:

1) Generate a statistically calibrated flow data set (3000 flows spread over six equal-numbered varieties).

2) Extracting the nine properties shown in Table 2 of each flow.

3) Coding items, data division (75% training / 25% test) in a class manner, and standardization of sensitive models.

4) Decision Tree, Random Forest, SVM, KNN, with the parameters described in Section 3.4.

5) Evaluation of each model on the independent testing group, as well as the implementation of the 5-fold cross verification on training data.

6) Confusion Matrix generates the best performing model, and a graphic importance graph of the random jungle model.

The total training data volume was 2250 flows, and the test data was 750 flows, with approximately 125 test flows per item, ensuring sufficient representation for each category when calculating the evaluation criteria. The design of the reproducibility was taken into account in the design of the experiment, by installing the random seed seed = 42, in all stages of data generation, training and division.

5. Results and Discussion

5.1 Comparative Performance Results

Table 3 presents the quantitative assessment results of all four models on the independent testing group, along with the average and cross-verification accuracy (5-fold) on training data.

model	Accuracy	Precision	Recall	F1-Score	Cross-validation accuracy CV (5-fold)
Decision Tree	97.07%	97.08%	97.07%	97.07%	95.11% ± 0.58%
Random Forest	97.87%	97.92%	97.87%	97.87%	97.69% ± 0.78%
SVM (RBF Kernel)	94.40%	94.44%	94.40%	94.40%	94.58% ± 0.71%
KNN (k = 7)	90.13%	90.31%	90.13%	90.09%	90.89% ± 0.96%

Academia Open

Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

Table 3: Compare the performance of the four classification models on the test group

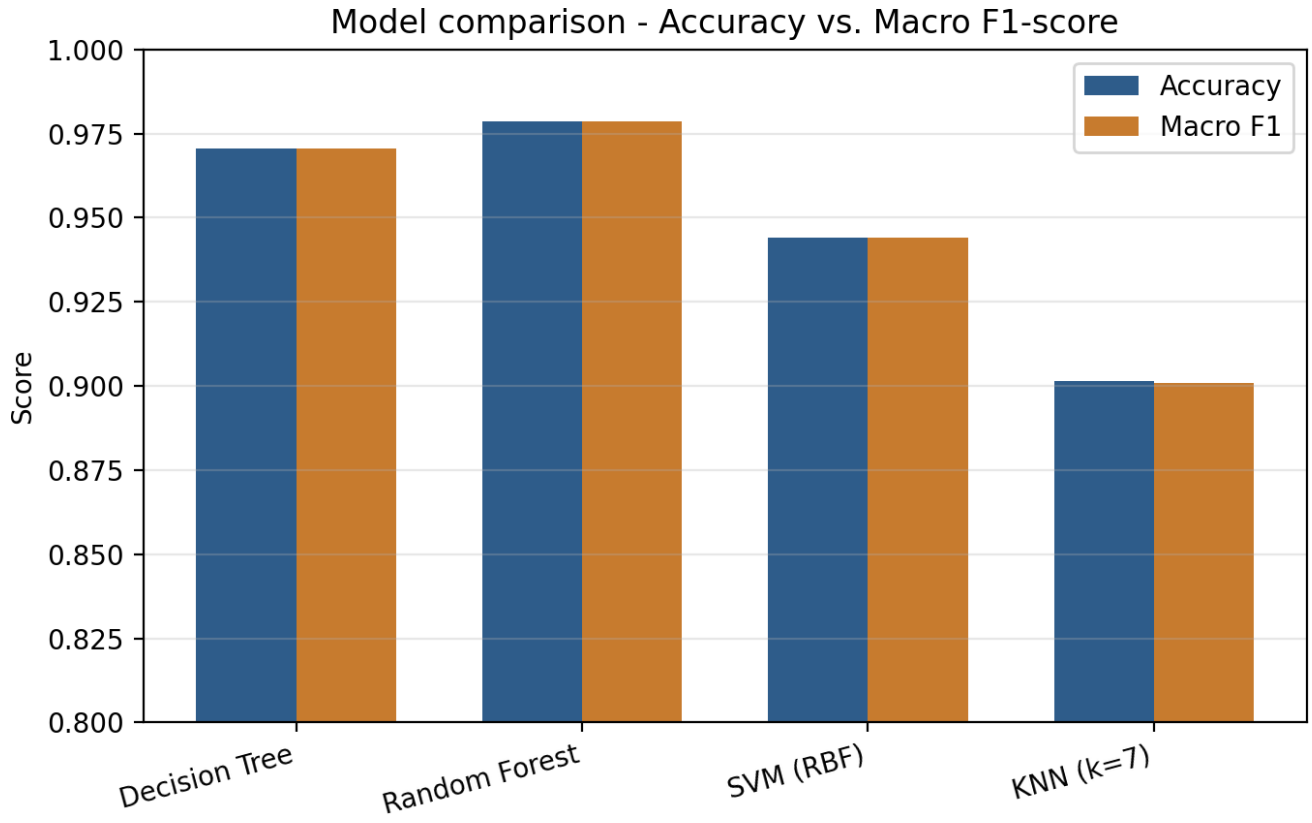

Figure 1.

Figure 1: A visual comparison between models in terms of overall accuracy and F1 score.

The results show that the Random Forest has achieved the best overall performance with 97.87% and F1 score of 97.87%, with the highest cross-verification stability (97.69%±0.78%), which is what It corresponds to what Salau and Beyene [9] reported of the superiority of aggregate tree models over individual models in network movement rating tasks within the SDN environment. The decision tree model came in second with a resolution of 97.07%, but the decrease in cross-verification accuracy compared to the test accuracy (95.11% versus 97.07%) indicates a higher degree of variance and greater sensitivity to the training sample change. In the random jungle, it is theoretically expected behavior that the random forest is originally designed to reduce the variation of the individual tree through the grouping.

On the other hand, the SVM model recorded an average resolution of 94.40%, a good performance that reflects the ability of the radial core (RBF) to represent the boundaries of non-linear separation between the varieties, but it remained below the level of the tree models in this experiment. KNN's algorithm has the lowest accuracy among the four models (90.13%), most likely due to its extreme sensitivity to local overlap between data points near the boundaries of the separation between statistically similar varieties, such as VoIP. and Gaming, which share similar ranges of packages and UDP.

5.2 Analysis of Confusion Matrix

To understand the error patterns more accurately, the Confusion Matrix was extracted for the random forest model as the best performer, as shown in Figure 2.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

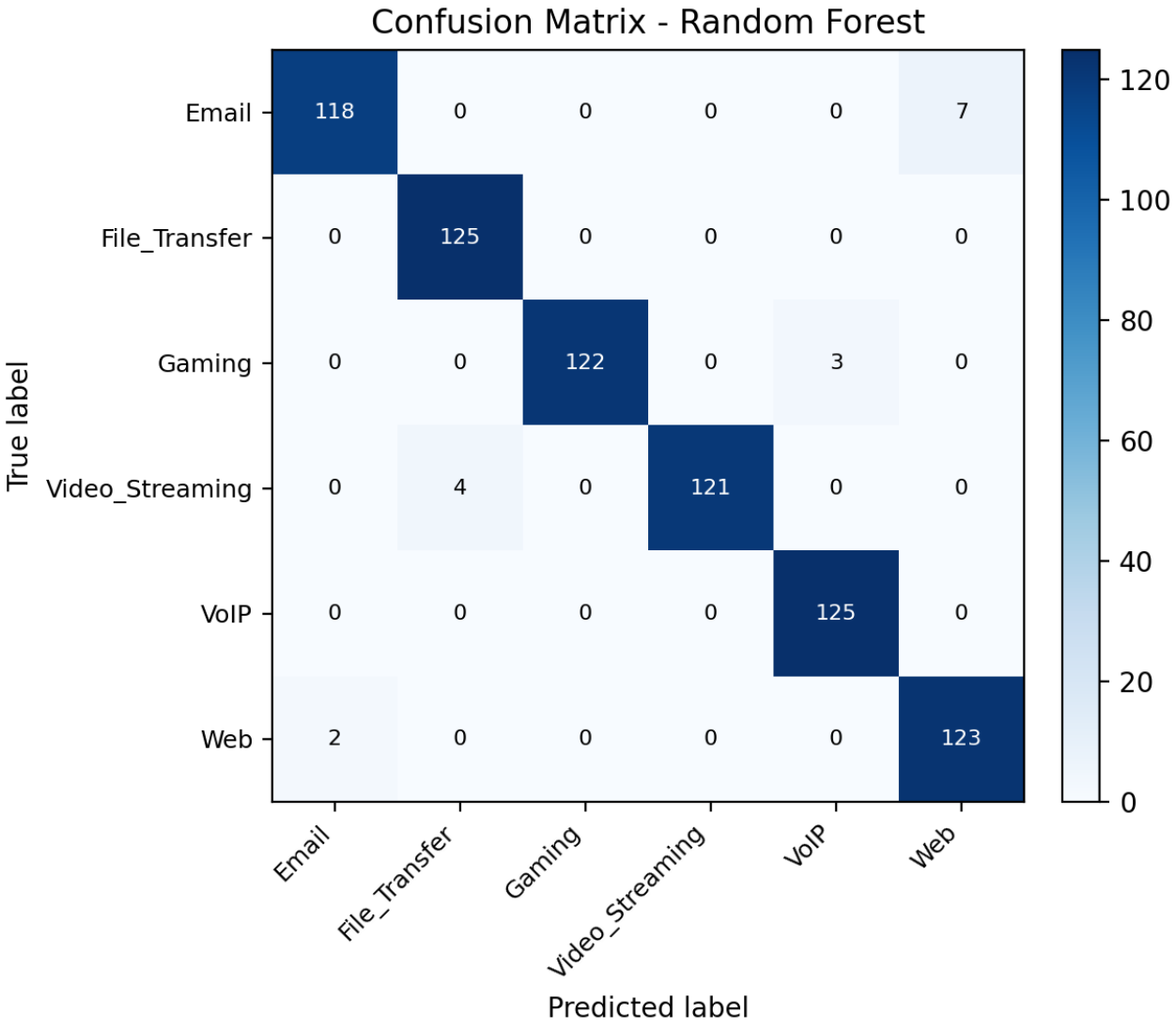

Figure 2.

Figure 2: Random Forest on the Test Collection

The matrix of confusion reveals that most classification errors are concentrated between VoIP and Gaming, which corresponds to the previous theoretical observation. Both classes are similar in their partial dependence on the UDP protocol and the relatively high rate of exchange of small beams in a short time, unlike the clearer differences between varieties such as File_Transfer. and Email (with small packages and longer intervals).

5.3 Feature Importance

Figure 3 shows the ranking of the nine properties according to their relative importance in the decision of the random forest model.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open
Vol. 12 No. 1 (2027): June
DOI: 10.21070/acopen.12.2027.15227

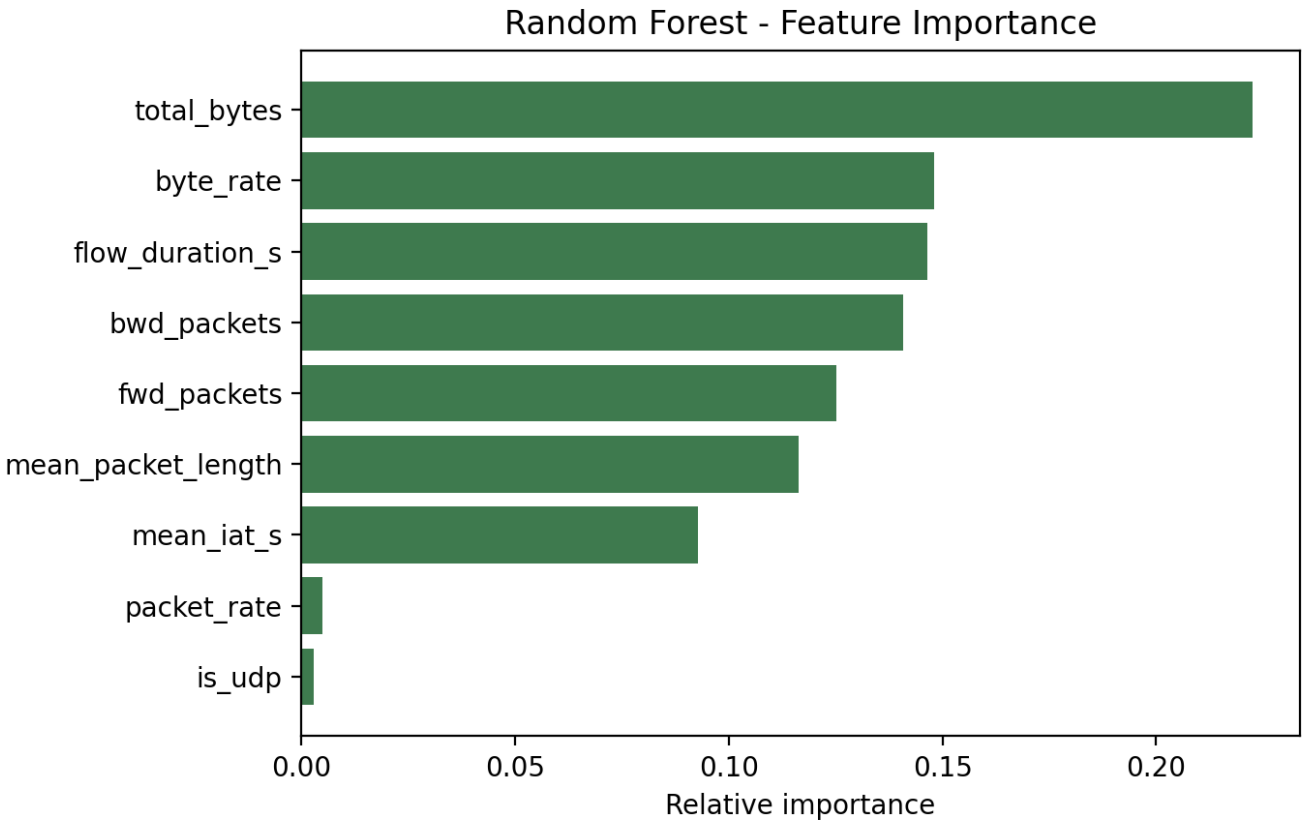

Figure 3.

Figure 3: The importance of relative characteristics according to the random forest model.

The results indicate that the byte_rate properties, the mean beam_packet_length, and the total_bytes are the most influential in the classification decision, which is a logical result. The networks are closely related to the nature of the application itself (e.g., its noticeable elevation in video broadcasting and the transfer of files versus its decline in e-mail and VoIP), which is consistent with the results reviewed by Lopez-Martin et al. [7] On the effectiveness of the time and volume characteristics in highlighting the networks of movement.

5.4 Comparison with Related Literature

The accuracy of the best model in this study (97.87%) falls within the range reported by recent studies to classify network movement based on statistical properties, which usually ranged between 90% and 99.5%, depending on the dataset, number of items and algorithm used. For example, Alshammari and Aleroud [14] recorded a resolution of 99.18% of the decision tree model on smart city network movement, while Elshewey and Osman [6] achieved a resolution of 99.49% using a compilation. A hierarchy of deep-encrypted HTTPS motion learning models, Antari et al. [4] reported 98.95% accuracy using the Transformer model on real institutional movement data. This comparison confirms that the findings in this research are consistent with the general trend documented in modern literature, supporting the reliability of the processing tube and the proposed classification.

5.5 Study Limitations

In spite of the positive results, this study involves a number of limitations that must be clearly recognized: first, rely on statistical calibration data rather than actual data captured from a real network or from certified reference datasets, due to the constraints of Internet access in an environment implementation; This calls for subsequent validation of results on realistic data. Second, the study is limited to only six types of data traffic, while real networks include dozens of various applications. Third, the failure to include deep learning models (such as bypass or repetitive networks) that have been proven in other studies [5] and [6] high competitiveness, especially with the entire encrypted data movement.

6.Conclusion and Future Work

This research provided an integrated applied study of network data traffic rating using machine learning techniques,

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

including an actual design and implementation of a complete data processing pipeline, from the extraction of nine statistical flow properties, to pre-processing. And ending with training and evaluation of four classification models (Decision Tree, Random Forest, SVM, KNN). The results showed that the random jungle model had a 97.87% resolution and a similar F1 score, with high cross-verification stability, while the KNN algorithm achieved the lowest relative performance due to its sensitivity to the overlap between the statistically convergent cultivars such as VoIP and Gaming.

These results confirm that the statistical properties extracted from the flow level constitute an effective and applicable alternative to categorizing the movement of data, even in the absence of the ability to read the package, which makes this approach particularly suitable for the environments of modern networks. It is dominated by encryption protocols. To complete this work in the future, the researcher recommends the following:

1)) Validation of results obtained on real reference datasets such as ISCX VPN-NonvPN [11] and CICIDS2017, to ensure the ability to generalize the movement of actual networks.

2)) Expanding the study to include a larger number of application varieties, including IoT and SDN.

3)) Explore deep learning models (CNN, LSTM, Transformer) and compare them with the classic models used in this research, especially in dealing with the entire encrypted data movement.

4)) Study the performance of the proposed models in a real-time/online classification, rather than the post-Full Classification.

References

1. Azab, M. Khasawneh, S. Alrabaee, K.-K. R. Choo, and M. Sarsour, "Network Traffic Classification: Techniques, Datasets, and Challenges," Digital Communications and Networks, vol. 8, no. 2, pp. 272–287, Apr. 2022.
2. M. Shen, K. Ye, X. Liu, L. Zhu, J. Kang, S. Yu, Q. Li, and K. Xu, "Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, vol. 25, pp. 791–824, 2023, doi: 10.1109/COMST.2022.3208196.
3. Y. Feng, J. Li, J. Mirkovic, C. Wu, C. Wang, H. Ren, J. Xu, and Y. Liu, "Unmasking the Internet: A Survey of Fine-Grained Network Traffic Analysis," IEEE Communications Surveys & Tutorials, Feb. 2025.
4. Antari, Y. Abo-Aisheh, J. Shamasneh, and H. I. Ashqar, "Network Traffic Classification Using Machine Learning, Transformer, and Large Language Models," arXiv preprint arXiv:2503.02141, 2025.
5. Z. Shi, N. Luktarhan et al., "BFCN: A Novel Classification Method of Encrypted Traffic Based on BERT and CNN," Electronics, vol. 12, no. 3, p. 516, 2023.
6. M. Elshewey and A. M. Osman, "Enhancing Encrypted HTTPS Traffic Classification Based on Stacked Deep Ensembles Models," Scientific Reports, vol. 15, Art. no. 35230, Oct. 2025, doi: 10.1038/s41598-025-21261-6.
7. M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey," Computer Communications, vol. 170, pp. 19–33, Mar. 2021.
8. R. H. Serag et al., "Machine-Learning-Based Traffic Classification in Software-Defined Networks," Electronics, vol. 13, no. 6, p. 1108, 2024, doi: 10.3390/electronics13061108.
9. O. Salau and M. M. Beyene, "Software Defined Networking Based Network Traffic Classification Using Machine Learning Techniques," Scientific Reports, vol. 14, no. 1, Art. no. 20060, Aug. 2024, doi: 10.1038/s41598-024-70983-6.
10. M. Sobh et al., "Software Defined Network Traffic Classification for QoS Optimization Using Machine Learning," Journal of Network and Systems Management, 2025.
11. E. Eslami and W. Hamouda, "Network Traffic Classification Using Self-Supervised Learning and Confident Learning," arXiv preprint arXiv:2509.23522, 2025.
12. M. Saqib, H. Elbiaze, and R. H. Glitho, "Adaptive In-Network Traffic Classifier: Bridging the Gap for Improved QoS by Minimizing Misclassification," IEEE Open Journal of the Communications Society, vol. 5, pp. 677–689, Jan. 2024.
13. Pekár, R. Plný, and K. Hynek, "Tutorial on Flow-Based Network Traffic Classification Using Machine Learning," arXiv preprint arXiv:2601.04089, 2026.
14. N. Alshammari and A. Aleroud, "A Comparative Study of Traffic Classification Techniques for Smart City Networks," Electronics, vol. 10, no. 15, p. 1854, 2021.
15. F. Rau, C. Herranz, I. Val, P. Georgieva, and J. Perez, "A Novel Flow-Based Online Network Traffic Classification Using Machine Learning," TechRxiv preprint, Aug. 2025, doi: 10.36227/techrxiv.175459507.75164419/v1.