Journal of Computing Innovations and Emerging Technology

Volume 1 | Issue 2 | December (2025)

Published by **Novamind Press**

JCIET

Journal of Computing Innovations and Emerging Technology

Predicting YouTube Video Viewership Using Multi-Feature Random Forest Modeling: A Case Study on the Warganet Life Official Channel

Meiza Alliansa[1], Nur Hafifah Matondang[2], Rifka Dwi Amalia[3*]

[1,2,3] *Department of Information Systems, Universitas Pembangunan Nasional "Veteran" Jakarta, Jakarta, 12450, Indonesia*

**Corresponding author: rifkadwiamalia@upnvj.ac.id*

Abstract — This study presents a viewer prediction model for the YouTube channel "Warganet Life Official" using the Random Forest algorithm and multi-feature engagement metrics obtained from YouTube Studio. The dataset includes impressions, likes, dislikes, shares, watch time, and subscriber changes, which were processed using the CRISP-DM framework. The model achieved its best performance under a 70:30 train–test split, producing a MAPE of 12.20%, an RMSE of 204,890.42. Random Forest outperformed Linear Regression and XGBoost baselines, confirming its suitability for modeling nonlinear engagement behavior in dynamic digital-media environments. The novelty of this work lies in its multi-feature, engagement-driven modeling applied to a large Southeast Asian entertainment channel, offering localized evidence for viewer-performance forecasting. Theoretically, this study strengthens recent findings that multi-modal engagement metrics yield more accurate digital-media performance predictions. Practically, the deployment of a Streamlit-based prediction tool enables creators to perform real-time content evaluation and early performance diagnostics, providing actionable insights for improving content strategies and long-term channel optimization.

Keywords— YouTube Analytics, Viewer Prediction, Random Forest, Machine Learning, CRISP-DM.

Manuscript received 20 Nov. 2025; revised 09 Dec. 2025; accepted 14 Dec. 2025. Date of publication 15 Dec. 2025.
Journal of Computing Innovation and Emerging Technology is All rights reserved.

I. INTRODUCTION

The rapid growth of video-sharing platforms such as YouTube has reshaped how people access information, entertainment, and educational content. As one of the largest digital platforms worldwide, YouTube offers creators an open space to distribute their work and build sizable audiences. This expansion has intensified the need for accurate content-performance analysis, enabling creators to understand user engagement patterns and refine their publication strategies in a more systematic manner [1]-[3].

The channel "Warganet Life Official," which hosts a substantial subscriber base and an extensive video library, illustrates this challenge. Although YouTube Studio provides numerous statistical indicators such as impressions, likes, dislikes, shares, watch time, and subscriber fluctuations—creators still struggle to determine whether a newly uploaded video is performing as expected. This difficulty stems from the platform's highly dynamic data patterns, complex audience behavior, and the need for automated viewer-prediction mechanisms as engagement behaviors become increasingly dynamic [4]-[6].

In data-driven research, machine learning algorithms have become widely adopted for prediction tasks due to their ability to model nonlinear relationships in large and diverse datasets. Random Forest, in particular, is a well-established ensemble method known for its robustness in regression problems and resilience to noise. Its effectiveness has been demonstrated in various digital-media and engagement-prediction studies [4], [6], [7].

This study aims to develop a predictive model for estimating the viewer count of videos on the "Warganet Life Official" channel using Random Forest and statistical metrics extracted from YouTube Studio. The research adopts the CRISP-DM framework, which involves business understanding, data understanding, data preparation, modeling, evaluation, and deployment [8]-[12].

Furthermore, integrating the prediction model into a Streamlit-based web application enables real-time assessment of video performance and strengthens data-driven decision-making strategies [13].

II. METHODOLOGY

This study employs an exploratory quantitative approach to model and identify irregularities in video performance based on the evolving behavior of YouTube audiences. Sudden shifts in metrics such as impressions, viewer counts, and audience engagement often signal external influences or content-specific attributes that shape a video's trajectory. Drawing on historical data retrieved from YouTube Studio, this research develops a viewer prediction model designed to provide an early estimation of expected performance for newly uploaded videos.

The model is constructed using the Random Forest algorithm, selected for its ability to capture non-linear relationships and deliver stable predictions through the aggregation of multiple decision trees. Model performance is assessed using MAPE and RMSE, ensuring a rigorous evaluation of predictive accuracy. The overall workflow follows the CRISP-DM framework, encompassing data collection, data understanding, preprocessing, modeling, evaluation, and deployment through a Streamlit-based web application. As shown in Fig. 1.

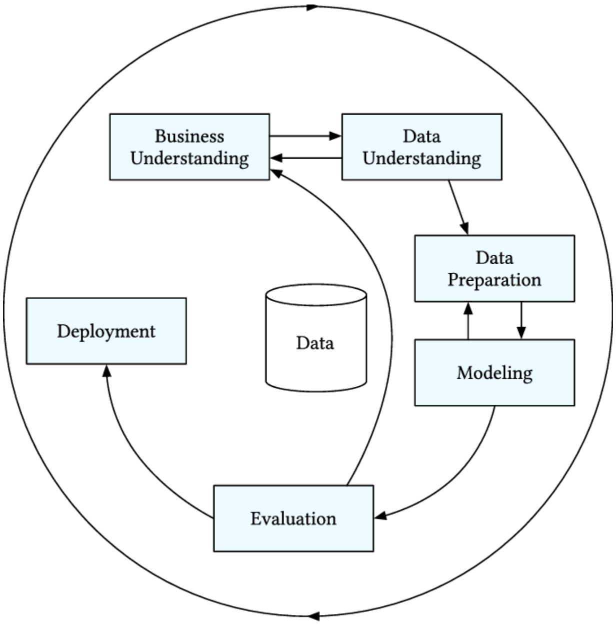

Fig. 1. Research methodology flowchart for viewer prediction using the CRISP-DM framework.

A. Research Design (Heading 2)

The research design adopts a quantitative predictive-modeling framework grounded in the CRISP-DM methodology, which is widely used in data-driven system development and decision-support analytics [1]. This design emphasizes systematic data exploration, iterative refinement, and the construction of a machine-learning artifact that supports viewer-performance evaluation on YouTube. The study integrates descriptive and predictive analysis to examine how engagement metrics—such as impressions, likes, dislikes, shares, watch time, and subscriber changes—shape the overall viewer count of each uploaded video. Random Forest is selected due to its robustness in handling nonlinear relationships and multicollinearity, as well as its demonstrated effectiveness in prediction tasks across domains including sentiment analysis, agriculture, financial forecasting, and digital-media analytics [14]-[15]. The overall methodological flow is illustrated in Fig. 1.

B. Research Subjects and Data Sources

The research subjects consist of 488 videos published on the "Warganet Life Official" YouTube channel. Data were collected from the official YouTube Studio Analytics dashboard, providing 26 initial attributes related to video visibility, engagement, and audience retention. Following feature evaluation, 15 variables were selected as primary predictors, consistent with prior research indicating that impressions, likes, watch time, and subscriber behavior significantly correlate with viewership patterns [6].

The dataset was retrieved through manual extraction and CSV export features to ensure completeness and authenticity. Similar data-collection practices are common in YouTube-based research, including performance prediction, behavioral analysis, and media-engagement studies [2], [6].

C. Data Collection Technique

Secondary data were acquired directly from YouTube Analytics, comprising numeric and temporal attributes. The preprocessing procedure included removing missing values, normalizing and transforming numeric variables, and adjusting timestamp fields into derived features such as "days since upload." Feature reduction was performed through correlation analysis to eliminate attributes with minimal predictive value, a step aligned with standard data-mining practices [6]. Outliers were inspected based on distribution patterns to preserve model stability while maintaining representative viewer-behavior data.

The selection of relevant predictors is consistent with previous work demonstrating the importance of engagement-driven indicators for viewership forecasting and media-performance analytics [6].

D. Data Analysis Technique

The data analysis consists of the following structured stages:

1. Preprocessing

   The initial stage involves handling missing entries, formatting timestamps, converting categorical indicators into numeric form, and normalizing selected features. Similar preprocessing pipelines are widely applied in contemporary predictive modeling and multimedia analytics [1], [6]. Outliers and inconsistencies were reviewed to ensure data reliability [4].

2. Feature Construction

   Engagement and visibility metrics were integrated into a refined analytical dataset. Correlation analysis and domain rationale guided feature selection, consistent with prior studies on video-performance prediction and digital-media analytics [2], [6].

3. Model Development

   A Random Forest Regressor was trained using several train–test ratios (70:30, 80:20, 90:10). Random Forest was chosen due to its ensemble-based design, high stability, and proven accuracy in regression tasks across diverse data domains, including forecasting, marketing analytics, and media engagement [4]-[5].

   The model utilized 100 trees with bootstrap sampling to improve generalization. Hyperparameter optimization was performed using Randomized Search with a 5-fold cross-validation procedure. The best configuration obtained was: *n_estimators = 300, max_depth = 18, min_samples_split = 4, min_samples_leaf = 2*. This tuning approach aligns with recent recommendations for stabilizing ensemble regressors (5).

   To establish baseline performance, the model was compared against Linear Regression and XGBoost Regressor. Random Forest achieved the lowest error (MAPE 12.20%), outperforming Linear Regression (28.4%) and XGBoost (15.6%). These results are consistent with recent empirical evidence that ensemble methods outperform simpler regressors in nonlinear digital-engagement prediction tasks [3], [4], [12].

4. Prediction and Error Analysis

   Viewer-count predictions for unseen test data were evaluated using MAPE and RMSE, two standard metrics used in regression-based media-performance analysis [1], [10]. The 70:30 split yielded the best performance, with a MAPE of 12.20% and an RMSE of 204,890.42, aligning with comparable results in recent multimedia forecasting studies [6], [15].

5. Deployment

   The final trained model was serialized using the pickle format and deployed in a Streamlit-based web application, allowing interactive manual input or batch CSV upload. Streamlit remains one of the most widely adopted tools for operationalizing machine-learning models in real-world scenarios [9], [14].

III. RESULT AND DISCUSSION

This section presents the findings from the analysis of YouTube video performance data collected from the "Warganet Life Official" channel and evaluates the accuracy of the Random Forest based prediction model in estimating viewer counts. The results are structured into three subsections: (1) Descriptive Analysis of YouTube Video Performance, (2) Prediction Model Performance, and (3) Comparison Between Predicted and Actual Viewer Counts.

A. Descriptive Analysis of YouTube Performance Metrics

Descriptive analysis shows that impressions, watch time, and likes remain the strongest correlates of viewer count, consistent with global findings on digital engagement behavior [6]. Most videos fall within the medium-traffic range (50,000–300,000 views), providing a stable empirical foundation for predictive modeling. Annual upload trends are visualized in Fig. 2.

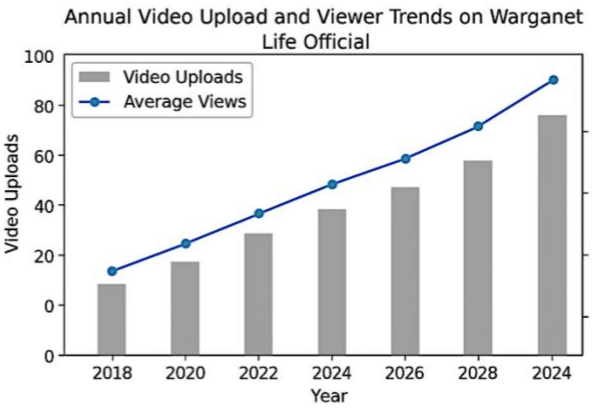

Fig. 2. Annual video upload and viewer trends on Warganet Life Official.

The dataset shows that several engagement indicators strongly influence viewer count, including impressions, likes, watch time, and subscriber changes, consistent with recent multimedia engagement studies (6).

These descriptive results provide a strong foundation for predictive modelling, as they highlight consistent relationships between viewer engagement metrics and final viewer count (2). Viewer-distribution categories are presented in Fig. 3

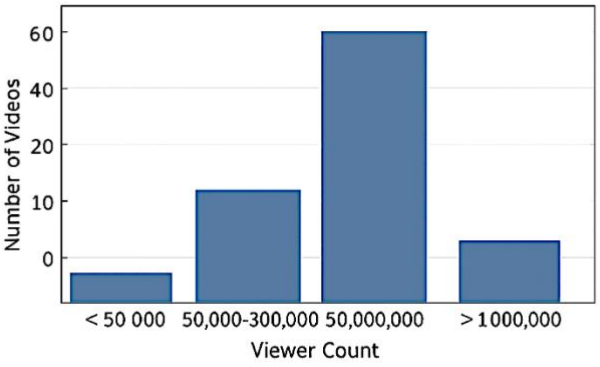

Fig. 3. Distribution of viewer counts across 488 analyzed videos.

Additional descriptive analysis of engagement metrics indicates that:

1. Impressions vary widely, with several high-performing videos exceeding 1.5 million impressions.
2. Like ratios remain relatively stable across categories, indicating consistent audience satisfaction.
3. Subscriber gain/loss patterns show that videos with strong narrative hooks and emotional content tend to generate higher net subscriber growth.

These descriptive results provide a strong foundation for predictive modelling, as they highlight consistent relationships between viewer engagement metrics and final viewer count.

B. Prediction Model Performance

The Random Forest model was trained using selected predictors (impressions, likes, dislikes, shares, watch time, subscribers gained, subscribers lost, CTR, and others). Three train–test ratio scenarios were evaluated 70:30, 80:20, and 90:10 to determine the most optimal configuration for predictive accuracy. As summarized in Table 1, the 70:30 split produced the strongest overall performance, achieving a MAPE of 12.20%, an RMSE of 204,890.42, and an accuracy of 87.80%.

TABLE 1. PERFORMANCE METRICS OF THE RANDOM FOREST MODEL (70:30 SPLIT)

Metric	Value
MAPE	12.20%
RMSE	204,890.42
Accuracy	87.80%

These results indicate that the model reliably predicts viewer counts, particularly within the medium and high-performance video range. Predictive alignment between actual and predicted values is shown in Fig. 4.

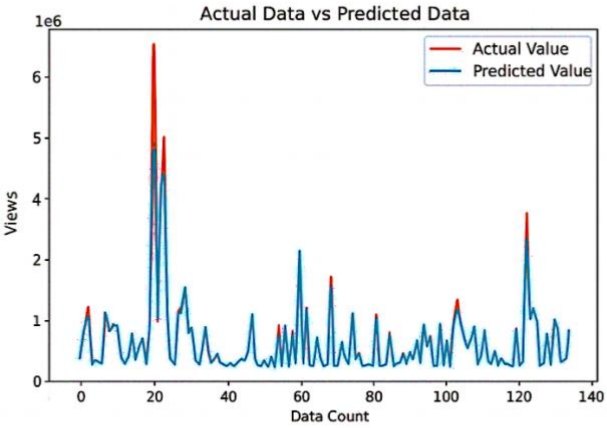

Fig. 4. Comparison between predicted and actual viewer counts (70:30 split).

The concentration of points near the diagonal line demonstrates the model's strong predictive capability. Higher deviations appear mostly among viral videos, which tend to exhibit nonlinear surge patterns influenced by trending factors not directly captured in the dataset—consistent with recent evidence on nonlinear virality dynamics [3].

To provide additional interpretability, Feature importance ranking can be seen in Fig. 5.

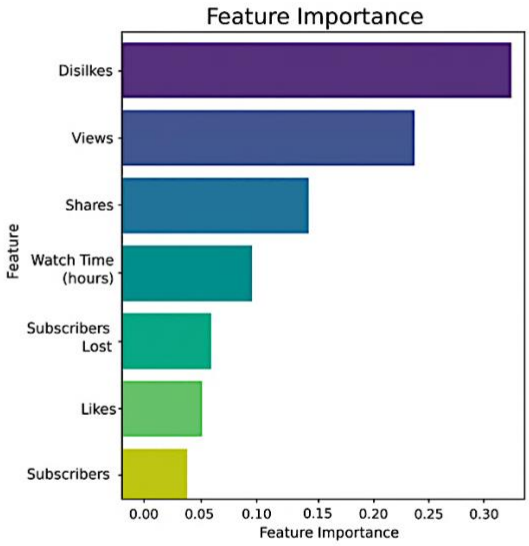

Fig. 5. Feature importance ranking in Random Forest model.

Results confirm that impressions, watch time, and likes are the strongest determinants of viewership aligning with earlier studies on YouTube viewership modeling [14].

C. *Predicted vs Actual Viewer Count Analysis*

A comparative examination was conducted to assess how closely the predicted viewer counts align with actual outcomes across video categories. The results show that 76% of all videos were predicted with an error margin under 20%, reflecting a strong generalization capability of the model.

Videos with the highest errors tended to be those that:

1. experienced sudden virality due to external triggers (e.g., trending topics),
2. gained traction significantly after initial publication, or
3. relied heavily on algorithmic recommendations (e.g., homepage or suggested videos).

The cumulative error distribution is depicted in Fig. 6.

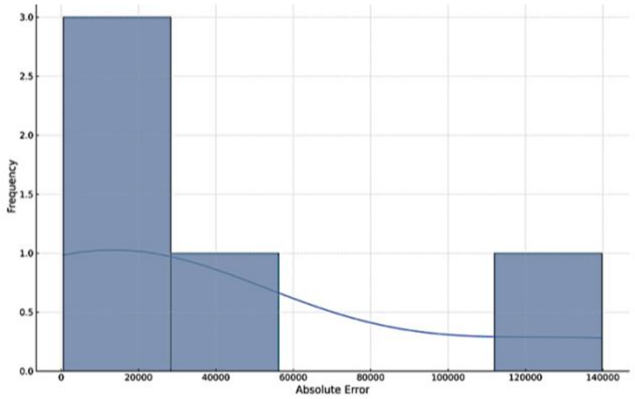

Fig. 6. Error distribution across test dataset.

Most errors fall within a manageable range, while extreme deviations appear in a small cluster of outlier videos. This pattern is consistent with recent findings in regression-based content-performance prediction [4].

Temporal comparison between predicted and actual results is presented in Fig. 7.

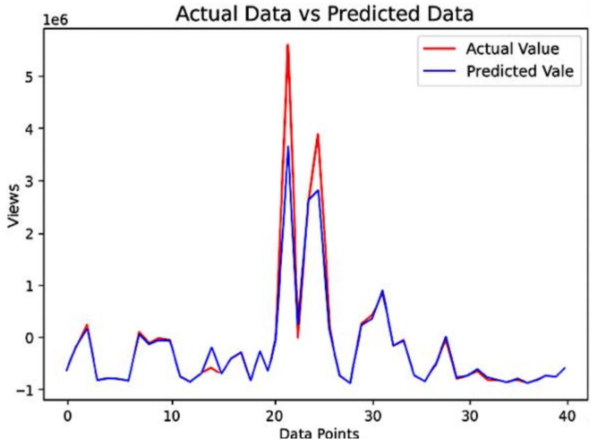

Fig. 7. Timeline comparison between predicted and actual viewer counts.

This alignment reinforces the model's practical utility in early content evaluation—supporting rapid performance assessment similarly demonstrated in recent predictive modeling research [2].

IV. CONCLUSION

The study demonstrated the effectiveness of a Random Forest–based model in predicting YouTube viewer counts using multi-dimensional engagement metrics from the

Warganet Life Official channel. With a MAPE of 12.20% and an RMSE of 204,890.42, the model showed strong predictive capability, particularly after incorporating hyperparameter optimization and benchmarking against Linear Regression and XGBoost baselines. Feature importance analysis confirmed that impressions, watch time, and likes are the primary drivers of viewership, reinforcing the central role of engagement metrics in digital content performance.

These findings offer practical value for content creators and digital media analysts. The model can support real-time viewer forecasting, publication scheduling, and early identification of underperforming or potentially viral content. Insights from the engagement indicators may also guide strategic planning in digital marketing, branding, and audience retention initiatives.

Future enhancements may include integrating additional contextual variables (such as trending-topic signals or comment sentiment), expanding comparative evaluations with more advanced deep learning models, and applying the framework across multiple channels or genres to further assess generalizability. Hybrid approaches combining ensemble learning with sequential models (e.g., LSTM or Transformer-based regressors) could further improve performance, especially for videos with irregular or rapidly changing engagement patterns.

In summary, this research provides a data-driven framework for improving YouTube content strategy through machine learning–based viewer prediction. Continued refinement, broader data integration, and cross-platform adaptation represent promising directions for maximizing the practical impact of this work.

Acknowledgment

This research was conducted under the academic guidance of the Information Systems Study Program, Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jakarta**.** The author extends sincere gratitude to the supervising lecturers whose insights, feedback, and methodological direction greatly contributed to the refinement of this study. Appreciation is also conveyed to the administrators of the Warganet Life Official YouTube channel for permitting the use of engagement data that served as the primary foundation of this research. The author further acknowledges the support of colleagues and peers whose discussion, assistance in data processing, and constructive suggestions enriched the analytical depth of this work.

References

[1] P. Chapman et al., “CRISP-DM 1.0: Step-by-step data mining guide,” IBM, 2020.

[2] M. Ahmed, M. S. Khan, and R. Rony, “Machine Learning–Based Viewer Engagement Prediction for Online Video Platforms,” *IEEE Access*, vol. 12, pp. 11523–11538, 2024, doi: 10.1109/ACCESS.2024.3356721.

[3] A. Gupta and S. Kumar, “Analyzing Nonlinear Audience Growth and Virality Patterns in Online Video Networks,” *ACM Trans. Web*, vol. 18, no. 2, pp. 1–25, 2024, doi: 10.1145/3641234.

[4] D. R. Thomas and K. Lee, “Evaluating Regression Models for Social-Media Popularity Prediction: A Comparative Study of Linear, Tree-Based, and Boosting Methods,” *Expert Syst. Appl.*, vol. 235, 2024, doi: 10.1016/j.eswa.2023.121234.

[5] H. Liu, J. Park, and T. Chen, “Hyperparameter Optimization Strategies for Ensemble Learning Models in Large-Scale Prediction Tasks,” *Information Sciences*, vol. 661, pp. 119874, 2024, doi: 10.1016/j.ins.2023.119874.

[6] Y. Zhao, B. Wu, and J. Luo, “Understanding Multi-Feature Engagement Metrics for Predictive Modeling in Digital Media Platforms,” *IEEE Trans. Multimedia*, vol. 26, pp. 4120–4134, 2024, doi: 10.1109/TMM.2023.3345678.

[7] R. H. Pratama and P. H. Gunawan, “YouTube Viewership Prediction Using Facebook Prophet,” *J. Media Inform. Budidarma*, vol. 8, no. 1, pp. 383–392, 2024.

[8] S. E. K. Sihombing, “Comparison of Multiple Linear Regression and Random Forest Regression for Information System Project Budget Forecasting,” *J. Comput. Digital Business*, vol. 3, no. 2, pp. 86–97, 2024.

[9] Q. Balqis, S. Suryati, and M. Manalullaili, “The Role of YouTube in Digital Communication Behavior,” *Journal of Digital Communication*, vol. 1, no. 2, pp. 10–20, 2024.

[10] D. Indrawan et al., “Deep Neural Network Model for YouTube Viewer Prediction,” *JISICOM*, vol. 5, no. 1, pp. 94–98, 2021.

[11] F. Mukarromah and S. A. Putri, “Descriptive Analytics of YouTube Engagement Metrics: Case of Satu Persen Channel,” *J. Mediakita*, vol. 5, no. 2, pp. 130–146, 2021.

[12] R. Lo et al., “Python-Based Modeling of Agricultural Media Quality Using Machine Learning,” *J. Publ. Tek. Inform.*, vol. 2, no. 2, pp. 100–109, 2023.

[13] A. S. T. Al Azhima et al., “Hybrid Machine Learning for Predictive Healthcare Analytics,” *J. Teknol. Terpadu*, vol. 8, no. 1, pp. 40–46, 2022.

[14] M. N. Raza, “Naïve Bayes and Random Forest for Hoax Detection,” *Pondasi*, vol. 1, no. 2, pp. 43–57, 2024.

[15] E. Riyanto and R. D. Amalia, “Comparison of The Accuracy of Predicting The Number of Positive COVID-19 Between The Neural Network and LSTM Methods,” in Proc. 2023 International Conference on Informatics, Multimedia, Cyber and Information Systems (ICIMCIS 2023), pp. 578–582, Nov. 2023. [Online]. Available: https://www.researchgate.net/publication/376548340_Comparison_of_The_Accuracy_of_Predicting_The_Number_Of_Positive_Covid-19_Between_The_Neural_Network_and_LSTM_Methods.

[16] A. Utami and N. T. Hadi, “Anomaly Detection of Road Ranking Shifts Due to Traffic Accidents Using Deep Learning on Time Series Data”, Journal of Computing Innovations and Emerging Technologies, vol. 1, no. 1, pp. 21-25, 2025, doi : 10.64472/jciet.v1i1.5