A Statistical and Machine Learning Analysis of the Significant Features of PPPoE Sessions for Quality Monitoring

Ayan Zhunussov
Department of Telecommunication Engineering, Almaty University of Power Engineering and Telecommunications, Almaty, Kazakhstan
jarmale@mail.ru

Alimzhan Baikenov
Department of Telecommunication Engineering, Almaty University of Power Engineering and Telecommunications, Almaty, Kazakhstan
a.baikenov@aues.kz

Tansaule Serikov
Department of Electronics and Telecommunication, S. Seifullin Kazakh Agro Technical Research University, Astana, Kazakhstan
tansaule_s@mail.ru

Olga Abramkina
Department of Cybersecurity, Almaty University of Power Engineering and Telecommunications, Almaty, Kazakhstan | Department of Cybersecurity, International Information Technology University, Almaty, Kazakhstan
olga.manank@gmail.com

Yelizaveta Vitulyova
National Scientific Laboratory for the Collective Use of Information and Space Technologies (NSLC IST), Satbayev University, Almaty, Kazakhstan | JSC "Institute of Digital Engineering and Technology," Almaty, Kazakhstan
lizavita@list.ru (corresponding author)

Received: 14 June 2025 | Revised: 13 July 2025 | Accepted: 16 July 2025

 DOI: https://doi.org/10.48084/etasr.12714

ABSTRACT

The present work explores the development and application of the method of indirect monitoring of telecommunication network quality based on the analysis of Point-to-Point Protocol over Ethernet (PPPoE) session parameters using machine learning methods as a key indicator of the network failures, the use of the K coefficient is justified based on the dynamics of PPPoE Active Discovery Termination (PADT) packets and the number of active PPPoE sessions. The paper describes the stages of data collection and preprocessing, including the conversion of session indicators from a "wide" format to a "long" format for ease of analysis. A statistical analysis of the significance of attributes (Analysis of Variance (ANOVA)-test, correlation analysis) was carried out, based on which a limited set of informative parameters of PPPoE-sessions (e.g., connection duration, frequency of disconnections, volume of transmitted data, connection establishment time) was selected. Linear Regression, Ridge, Lasso, Random Forest, and Support Vector Regression (SVR) models were trained and comparatively evaluated on these attributes to predict the *K* value. The symbolic regression experiment provided an analytical formula to confirm the correctness of the selected *K* value. The comparative analysis by the Mean Squared Error (MSE) and Coefficient of Determination (R^2) metrics showed the advantage of Random Forest model ($R^2 \approx 0.90$, MSE ≈ 0.0001), which indicates the high efficiency of the proposed approach. The significance of the study lies in

demonstrating the possibility of the early detection of the network quality anomalies without a direct analysis of the traffic content, which increases the efficiency of monitoring the quality of telecommunication services.

Keywords-machine learning; PPPoE; Quality of Service (QoS); statistical analysis; network monitoring; broadband networks

I. INTRODUCTION

The current state of telecommunication networks is characterized by a rapid increase in the number of subscribers and the volume of transmitted data, which necessitates the implementation of effective approaches to the service quality monitoring and management [1–3]. A key parameter in evaluating the service quality is the analysis of the network structure, which enables a timely detection of the deviations in the network performance and ensures the required Quality of Service (QoS). Particularly important in this context is the assessment of parameters related to the PPPoE transport protocol (Point-to-Point Protocol over Ethernet). Approaches to modeling integrated quality assessment systems in cybersecurity contexts can also inform the network QoS evaluation frameworks [4]. This protocol is widely used to provide user access to network resources over broadband connections and is actively employed in the infrastructure of the internet service providers [5]. The increasing complexity of broadband networks and the growing number of subscribers require advanced monitoring techniques to ensure consistent QoS. Machine learning approaches have shown significant promise in addressing these challenges by enabling predictive and automated QoS management in telecommunication networks [6].

The direct monitoring of all infrastructure components can be technically complex and computationally expensive, especially under high traffic dynamics. Often, direct monitoring relies on the content of higher-level TCP/IP protocol packets. However, in the presence of tunneling technologies or encrypted data, the direct access to payloads may be restricted. Consequently, there is growing interest in the indirect monitoring methods that do not require access to packet contents, but instead rely on behavioral traffic indicators and PPPoE session statistics [7, 8]. PPPoE is a widely adopted protocol for managing the user sessions in broadband access networks. Its integration into modern IP network architectures, such as those leveraging Asterisk PBX, ensures reliable connectivity and supports QoS monitoring [9]. Analyzing session characteristics, such as the connection duration, data volume, session drops, and establishment latency, allows for an objective assessment of the user experience quality and timely response to emerging issues. These indicators can also serve as input features for machine learning algorithms to detect abnormal behavior and anomalies, which may indicate network failures, bandwidth congestion, authentication errors, DDoS attacks, or routing issues.

This study introduces an approach to indirect monitoring of service quality based on the behavioral analysis of PPPoE sessions, without accessing payload data. The proposed anomaly detection framework uses parameters such as session duration, disconnection rates, traffic volume, and session initiation delays. A comparative analysis of machine learning models-including Linear Regression, Ridge, Lasso, Random Forest, and SVR-is applied to predict the failures and QoS violations based on PPPoE-specific session features. The validation on real access logs from a major telecom provider confirms the approach's practicality and applicability.

The rationale and prospects for the proposed method are supported by prior research. The use of PPPoE packet statistics for QoS monitoring has been investigated [10], as well as machine learning applications in predicting the network anomalies [6]. These efforts laid the groundwork for assessing the feasibility of indirect network diagnostics using a metric K, calculated from the dynamics of PADT packets and the number of active sessions.

In [11], tasks such as classification, optimization, and network security are explored, along with challenges, like data imbalance and real-time processing constraints. On this basis, the examination of PPPoE-specific features for QoS monitoring emerges as both relevant and novel. Research on fault detection in Industrial IoT (IIoT) environments has demonstrated the potential of machine learning models for identifying failures based on device state characteristics, an approach that could be adapted to broadband access systems like PPPoE. A system-level approach to the flow optimization in multiservice environments has been demonstrated in [12], highlighting the importance of considering routing topology and resource constraints in telecommunication service quality management [12].

AI-driven monitoring systems have also been investigated in [13], where the integration of machine learning is highlighted for its role in anomaly detection and failure prediction in network environments. Ensemble approaches have shown effectiveness in heterogeneous network conditions, including LiFi and RF handovers [14]. For example, authors in [15] demonstrate the use of support vector machines and time-series forecasting for failure prediction in optical networks.

Furthermore, large-scale comparative studies of anomaly detection techniques in time-series data have been conducted in [16], assessing both traditional statistical and deep neural methods. These have been deployed in real-world systems, such as the Microsoft telemetry anomaly detection service described in [17], which relies on hybrid machine learning models.

Advanced architectures, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have proven effective in detecting anomalies. The CNN-based DeepAnT approach [1] offers high accuracy without requiring labeled datasets, while the LSTM models have shown strong performance in analyzing the temporal patterns in network data [2].

Nonetheless, there is a notable gap in applying interpretable machine learning to session-based metrics in PPPoE environments. While symbolic regression is emerging as a method to generate human-readable, analytical models from data, its application in PPPoE traffic analysis remains limited. This study fills this gap by introducing and validating a novel coefficient K for session instability detection, derived from measurable PPPoE parameters and verified through regression modeling and symbolic analysis.

II. MATERIALS AND METHODS

The methodological framework of this study involves a sequential implementation of several key stages aimed at constructing a model for evaluating QoS based on the analysis of PPPoE traffic under real operational conditions.

A. Data Collection

As an initial stage of the methodology, a structured data collection system was deployed to extract PPPoE session statistics from a real broadband network environment. Session log files and aggregated performance metrics collected at the Broadband Remote Access Server (BRAS) level were used as data sources. This configuration allowed capturing key metrics, such as VLAN IDs, IP interfaces, session durations, traffic volumes, and disconnection rates [18]. Previous simulation-based studies using tools, such as Cisco Packet Tracer, have shown the effect of the ICMP packet sizes on latency metrics, relevant to the session stability assessment [19].

B. Quality Indicator (K) Formation

Based on the collected statistics, an integral indicator – referred to as coefficient K – was calculated to reflect the stability and reliability of the data transmission across different directions. This coefficient aggregates characteristics, such as the number of disconnections, reconnections, and packet losses recorded on access interfaces and virtual network segments [20]. The following initial attribute-factors are considered to predict the probability of PPPoE session loss:

- PADI - number of connection initiations
- PADO - number of connection offers from the server
- PADR - number of connection requests from the client
- PADS - number of confirmed sessions
- PADT - number of terminated (broken) sessions
- Summary sessions - total number of active PPPoE sessions in the network
- Time - time (e.g. hourly index) reflecting the long-term trend

It was necessary to determine which of these attributes have the most significant impact on PADT (session loss) to find the key factors that explain the variation in the number of PPPoE session breaks. The task was solved by methods of statistical data analysis. First the correlation coefficients between the attributes and the target variable were calculated, and ANOVA with Fisher's F-test and p-value estimation was performed to test the significance of the factors. Also, the statistical significance of a trait was defined as the ability of its variation to explain the variation in response (session losses) at a level that exceeds random noise with low confidence level ($p < 0.05$). The study analyzed the time series from the obtained statistics. The trait $K = (PADT_t - PADT_{t-1})/S_{t-1}$ was introduced as a target trait reflecting the proportion of gaps per load interval. Additionally, 10 static attributes $Ki,j = (Xi - Xj)/S$, where Xi, Xj are taken from PADI, PADO, PADR, PADS, and PADT were generated. In order to assess the significance of the features, the ANOVA F-test was applied and correlation coefficients were calculated with K [19].

Thus, the goal is to identify the key factors (both static and dynamic) that determine the probability of the PPPoE session rupture and confirm their significance using statistical methods. The results of this analysis allowed to narrow down the set of attributes for subsequent modeling and machine learning, improving the interpretability and effectiveness of the network failure prediction model [6].

C. Prediction Using Machine Learning Models

This stage involves the task of regression-based prediction of the K coefficient using network metrics. Predictive features included quantitative characteristics of traffic and user activity. Both classical regression methods (Linear, Ridge, and Lasso regression) and more advanced models (Random Forest and SVR) were employed for the analysis [23, 24].

D. Comparative Model Evaluation

This stage involves an experiment aimed at evaluating the accuracy and robustness of the models under real network load conditions. Performance metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R^2 were used to assess the model quality. The most effective model was selected based on its accuracy in approximating the actual K value and its suitability for real-time QoS monitoring in telecommunication systems [25]. The statistical data collection followed the methodologies proposed in [6, 10, 14]. The initial dataset is presented in Table I, where the column values correspond to different network directions (interfaces such as *xe-...*, *ps...*, etc.), and rows represent the time stamped snapshots (a total of 98-time intervals were recorded). For each direction, the following indicators were recorded: Summary Sessions, PADI, PADO, PADR, PADS, and PADT [25].

TABLE I. PPPoE PACKET STATISTICS FOR MULTIPLE DESTINATIONS WITH TIMESTAMPS

Time	xe-0/2/0.3221828270					
	Summary sessions	PADI	PADO	PADR	PADS	PADT
2024-05-23 14:54:20 ALMT	340	0	27610	0	18706	17001
2024-05-23 14:59:31 ALMT	340	0	27613	0	18709	17004
2024-05-23 15:04:47 ALMT	340	0	27615	0	18711	17006
2024-05-23 15:41:39 ALMT	339	0	27650	0	18741	17037

In Table I, the data are organized in a two-level header format, where the first row defines the general categories and the second row specifies the parameters within each category. This format complicates the analysis, since each time slice is presented in a single row, and different metrics are located in separate columns. For ease of processing, the table was loaded into Python using pandas, with the headers read from header = (0,1), which allowed for the correct interpretation of the data hierarchy. Next, a transformation to long format was applied in which the columns with parameters were reduced to two key attributes: "Parameter name" and "Value", which allowed for a unified data structure. This transformation simplifies aggregation, regression analysis, and subsequent model processing. Then, within each "Direction" group (i.e., a specific interface), timestamps were sorted and the difference Δ(PADT) was calculated. Based on this difference and summary sessions, the *K* coefficient was calculated as [5, 7]:

$$K = \left(PADT_i - PADT_{(i-1)}\right)/S \quad (1)$$

where $PADT_i$ and $PADT_{i-1}$ are the values of the PADT counter (number of session terminations) at the current and previous time points, and *S* is the total number of active sessions [6].

This formula represents the proportion of sessions terminated within the interval between observations $i-1$ and i. The higher the value of *K* is, the more abrupt is the increase in the session terminations during that time window. Thus, the coefficient *K* serves as an indicator of abnormal disconnection spikes. Under normal network operation, *K* remains low, whereas during major failures (e.g., simultaneous disconnection of many PPPoE sessions), *K* approaches 1 or exceeds the typical threshold values. This metric enables real-time detection of anomalies associated with connection instability.

It is important to note that in addition to specific metrics, like *K*, modern network administration increasingly employs machine learning methods for anomaly detection. However, compared to these complex approaches, the *K* metric offers a simpler and more interpretable indicator specific to PPPoE sessions.

To predict the value of *K* based on PPPoE session parameters, several regression models are employed. Each model aims to minimize the error between the predicted and actual *K* values (typically measured using MSE) and is based on specific assumptions about the data structure.

The following algorithms are used in this study:

- Multiple Linear Regression

This model assumes a linear relationship of the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (2)$$

where *x* represents the features (e.g., counts of PADI, PADO, PADR, PADS packets, session count *S*, etc.), and β are the model coefficients. The model is trained using the least squares method, which minimizes the sum of the squared deviations from the actual values (the MSE loss function). Linear regression is a baseline approach for estimating *K* and provides interpretable coefficients showing the impact of each feature [5]. However, it cannot capture the complex non-linear effects.

- Ridge Regression

This is a modified linear model that incorporates the L2 regularization. A penalty term $\alpha\sum_j \beta_j^2$ is added to the loss function to prevent overfitting and address the multicollinearity among features [26]. While the model remains linear, the penalty reduces the variance of the coefficient estimates. Ridge regression is particularly useful when many correlated PPPoE metrics are present, as it smooths the influence of less informative features.

- Lasso Regression

Another regularized linear model, Lasso applies an *L1* norm penalty $\alpha j\sum \mid \beta j \mid$. This penalty drives some coefficients to zero, effectively performing feature selection [27]. This means that the model automatically identifies the most influential parameters (e.g., it may exclude redundant metrics that have little impact on *K*). Lasso regression is useful when selecting a subset of meaningful indicators from a wide array of network statistics. The loss function is also based on MSE but includes an *L1* penalty; solving it requires specialized optimization methods due to its non-smooth nature, but modern libraries offer efficient implementations.

- Random Forest Regression

This is an ensemble method consisting of multiple decision trees trained on different data subsets and feature combinations [25]. Each tree constructs recursive if-then rules, partitioning the feature space to minimize MSE within the leaves. The final prediction is the average output of all trees, which reduces variance (the bagging effect). Random Forest can model the complex non-linear relationships between the PPPoE metrics and the *K* coefficient, accounting for the interactions between features (e.g., the combined influence of PADI and PADO). The advantage of Random Forest lies in its high accuracy and robustness to noise, while its main drawback is the low interpretability—since the model may include hundreds of trees, direct analysis is difficult.

- Support Vector Regression

The Support Vector Machine (SVM) method has been extended to regression tasks through the introduction of the ϵ -insensitive zone. The objective of SVR is to approximate the dependency $y \approx f(x)$ with a function *f(x)* such that deviations $\mid y - f(x) \mid$ smaller than ϵ are not penalized, while larger deviations are minimized using the largest possible margin.

The final optimization problem minimizes the functional $\frac{1}{2}|w|^2 + C\sum_i max(0, |y_i - f(x_i)| - \epsilon)$, which results in a sparse solution – only a subset of data points (called support vectors) that is used to define the regression.

SVR with a kernel (such as the Radial Basis Function (RBF) used in this study) is capable of modeling the non-linear relationships by projecting the original features into a higher-dimensional feature space, and has been successfully applied in

previous works for predicting the user behavior in IPTV systems [29]. In the context of PPPoE, this allows the model to capture the complex interactions between indicators (e.g., the non-linear influence of session count *S* on the behavior of coefficient *K*). SVR typically requires the tuning of hyperparameters - ϵ width and the regularization parameter *C* to manage the trade-off between the training accuracy and generalization performance.

Each of the models described above was trained to predict the *K* coefficient based on features derived from PPPoE session data, including the number of PADI, PADO, PADR, PADS packets, PADT differences, and more. The training was conducted on historical session logs. The dataset was split into training and testing subsets, and the models were fitted on the training data by minimizing the MSE loss. The performance was then evaluated on the test data.

It is worth noting that the regularized models (Ridge and Lasso) help prevent overfitting to random fluctuations in metrics, which is particularly relevant for high-frequency traffic variations. Random Forest and SVR, due to their ability to model non-linearities, often outperform linear models in predicting the *K* value [28]. This observation was confirmed experimentally, as Random Forest and SVR typically achieved the highest R^2 values on the test data. However, these complex models generally require more data for reliable training and are more difficult to interpret.

To quantitatively assess the model performance, standard regression metrics were used: MSE and the R^2 [31]. MSE is computed as the average squared difference between the model's predictions $\widehat{y_i}$ and the actual values y_i:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (4)$$

This metric has the dimension of the square of the original value and is interpreted as the average degree of model error. The smaller the MSE is, the closer the predictions are to the real values (MSE = 0 means a perfect match). MSE is convenient for optimization (differentiable, single minimum) and, therefore, often acts as a loss function when training regression models. However, its value is difficult to interpret in absolute terms. It is important to compare the MSE of different models with each other on the same dataset. For a more interpretable quality assessment, the R^2 is used. It is defined as the proportion of response variance explained by the model:

$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{y_i})^2}{\sum_i (y_i - \bar{y})^2} \quad (5)$$

where $\bar{y}$ is the average value of the actual *y* based on the test data.

The R^2 value varies from 0 to 1 (for a perfect prediction, R^2 =1; for a model that does not outperform the trivial mean, R^2=0; a negative R^2 is also possible if the model is worse than a constant mean). Thus, R^2 shows the proportion of the variation in the metric *K* that can be explained by the model. A high R^2 indicates that the model captures the main pattern in the data. In network problems, where a significant portion of fluctuations can be random or caused by external factors, even moderate R^2 values (~0.5–0.7) are considered good. According to research, R^2 is often more informative for comparing models than absolute error values, and is preferable when assessing the quality of regression. In this study, the models were compared primarily by R^2 on the test sample, and MSE was used to understand the scale of errors. For example, the best of the classical models (Random Forest) showed an R^2 of about 0.85, which means that about 85% of the variation in the coefficient *K* is explained, with the mean square error deviation of about several units in absolute values of *K*. These metrics confirm the correctness of the model and allow a quantitative comparison of different algorithms.

In addition to the listed algorithms, symbolic regression implemented through the genetic programming method is of particular interest. Symbolic regression attempts to explicitly identify an analytical formula that relates the target variable (*K*) to the features (PPPoE metrics) through the evolutionary enumeration of various functions and feature combinations. In this case, the gplearn library (*SymbolicRegressor* class) is used, which operates on a population of candidate formulas and evolves them, similar to biological evolution, based on their fitness (accuracy).

The principle of symbolic regression lies in a random population of mathematical expressions built from a basic set of operations (+, −, *, /) and random constant starts. Each expression is a genetic program, usually represented as a tree (nodes are operations, leaves are variable-features or constants). For each candidate, the loss function is calculated, in this case, MSE on the training data. Then the most accurate formulas (with the lowest MSE) are selected from the population. These are subjected to genetic operations such as recombination (crossing), the exchange of parts of expressions between two formulas, and mutations, which involve random changes to operations or constants in the formula, among others. The algorithm parameters set, such as the population size (in the work, about 2000 expressions), the number of evolutionary generations (20-50), the probability of crossover (for example, 0.7) and mutations of different types (the general order is 0.2-0.3). After each generation, the newly obtained formulas are evaluated by the MSE metric, and the process is repeated, gradually improving the accuracy of the best expressions. The evolution stops either when a given number of generations is reached, or when a given stopping criterion is reached. For example, if the MSE falls below a given threshold (in this case,10^{-5}). As a result, symbolic regression tries to find an explicit formula for *K* based on PPPoE features. For instance, one of the obtained best formulas (conditionally) could look like:

$$K \approx 0.0025 \cdot PADT_{diff} - 0.0001 \cdot PADO + 0.05 \quad (6)$$

where $PADT_{diff} = PADT_i - PADT_{(i-1)}$ is the difference in PADT (already used in the definition of *K*), and *PADO* is the number of PADO packets per interval.

Equation (6) indicates that an increase in the number of session breaks ($PADT_{diff}$) directly increases *K*, while an increase in the server responses (*PADO*) slightly decreases *K* (e.g. due to a more stable session with a larger number of offers

from hubs). It is important to emphasize that when using symbolic regression, the same quality metric, MSE, was used as the objective function (fitness). This means that evolution directly optimized the mean square error of the prediction *K*. In addition, complexity limiting methods (e.g. tree depth penalty or maximum depth limit) are used to prevent the excessive growth of the formula sizes. gplearn implements the stopping criteria parameter, which allows stopping the evolution when the improvement becomes insignificant (MSE < 10^{-5}). After training all models (linear, Ridge, Lasso, Random Forest, SVR and symbolic), their qualities were compared using the MSE and R^2 metrics on the delayed test set.

III. RESULTS AND DISCUSSION

The processing of the "wide" table allowed to transform the data into approximately 1,666 rows in a "long" format. During the model training phase (80% of data for training, 20% for testing), the following results were obtained. Table II shows the calculated statistical indices for the original (non-derived) features with respect to the target variable PADT (number of broken PPPoE sessions) based on the accumulated data. The features are sorted by decreasing the *F*-value.

TABLE II. STATISTICAL INDICATORS FOR BASIC FEATURES (RELATIVE TO PADT)

Feature	*F*-value	*p*-value	Correlation with PADT
PADS	9.07×10^{4}	1.15×10^{-144}	+0.9995
Time (hours)	1.94×10^{4}	1.35×10^{-112}	+0.9975
PADO	7.42×10^{3}	1.01×10^{-92}	+0.9936
Summary session	1.29×10^{2}	1.79×10^{-19}	–0.7574
PADR	1.30×10	1.00×10^{-5}	+0.900
PADI	≈0	1.00	-

The PADS, Time and PADO features, show the highest *F*-values and negligible *p*-values (<0.001), i.e. they have a significant effect on the number of session breaks. PADI is statistically insignificant and can be excluded.

The significance of the derived features (*K* coefficient and differences). Next, the significance of the dynamic *K* coefficient and all combinatorial features of the form $(X_i - X_j)/S$ were analyzed. The target variable (K_t) for this analysis is the calculated for each interval (percentage of sessions lost per interval). For each candidate, $\frac{X_i(t)-X_j(t)}{S(t-1)}$ is evaluated to see how well it explains the variation in the true K_t. A univariate ANOVA *F*-test [22] for the regression $K_{(i,j)} \rightarrow K$ was applied, and the *p*-values and correlations with *K* were calculated [3]. The results for the most representative combinations are summarized in Table III.

Table III shows the striking superiority of the *K* coefficient over any other combinations. The reference feature *K* (the difference of PADT on the interval normalized by the session) gives a huge $F\approx1250$ at $p<10^{-200}$, while the closest "chaser" - the combination $(PADT-PADS)/S$ has an *F* of about 350. The contributions of the others are even smaller $(PADT-PADR)/S$ gives $F\approx180$, $(PADR-PADS)/S$ is of the order of 95, etc. Statistically insignificant results ($p > 0.05$) are observed for most combinations that do not contain PADT. For example, differences including only PADI, PADO, PADR (without PADT or PADS) have an *F* less than 1 and $p\approx0.3$-0.5, i.e., their effect on *K* is random. Even if some of such features are correlated with the modulo *K*, they do not contribute new information already accounted for by other factors.

Figure 1 shows the heat map of the correlation between all combinatorial features $K_{(i,j)}$ and the target *K*. It confirms the above quantitative findings. In the row/column corresponding to the reference *K*, the bright red cells are observed only for traits containing PADT and/or PADS (upper left block of the matrix).

These traits are highly correlated with *K* (coefficients $\rho \approx 0.8 - 0.99$). In contrast, the block of combinations without PADT/PADS (lower right corner of the matrix) has a pale coloration, indicating low correlations ($|\rho|$ close to 0). Thus, of all the combinatorial indicators, the ones that are statistically significant are those related to the immediate events of session termination (PADT) or confirmation (PADS), or their difference (e.g., PADO-PADS). This approach to identifying the most critical influencing parameters aligns with role-based analysis methods used to determine the functional stability of complex systems [29]. These results are consistent with past studies. For reliable accident prediction, dynamic attributes involving counters directly reflecting the session breakage are most informative, while arbitrary combinations of indirect metrics do not improve the model.

According to the results of correlation analysis and *F*-test, the most significant was the dynamic feature $K = (PADT_t - PADT_{t-1})/S_{t-1}$ ($F> 46000$, $p <$ 1e-200). Prior work has also proposed the use of statistical criteria, such as Pearson's chi-squared to assess the transmission quality over multiservice networks [27]. From the static features, *PADT - PADS)/S*, *PADT - PADO/S*, and *PADO - PADS/S* stood out. Pure counts of PADT, PADS, and PADO also showed a high correlation with *K* (corr > 0.84).

TABLE III. COMPARISON OF SIGNIFICANCE FOR DYNAMIC FEATURE K AND COMBINATIONS $(x_i\text{-}x_j)/S$

Feature	F-value	p-value	Correlation with K
$K = (PADT_i - PADT_{(i-1)})/S_{(t-1)}$	1.25×10^{3}	$<10^{-200}$	+0.998
(PADT – PADS)/S	3.50×10^{2}	2.3×10^{-80}	+0.950
(PADT – PADO)/S	2.73×10^{2}	1.1×10^{-4}	+0.982
(PADT – PADR)/S	1.80×10^{2}	1.1×10^{-35}	+0.890
(PADO – PADS)/S	1.20×10^{2}	1.0×10^{-25}	+0.900
(PADI – PADO)/S	0.80	0.37	+0.020

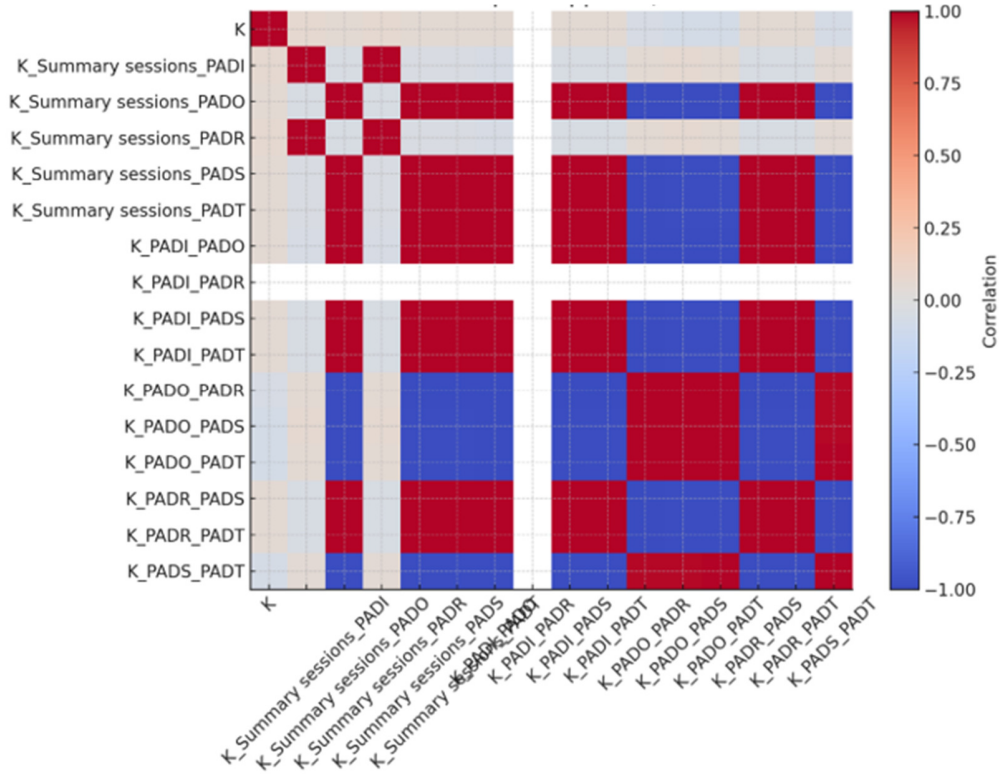

Fig. 1. Heat map of the correlation matrix for the coefficient K and all combinations $(X_i\text{-}X_j)$/S (the redder the cell, the higher the correlation).

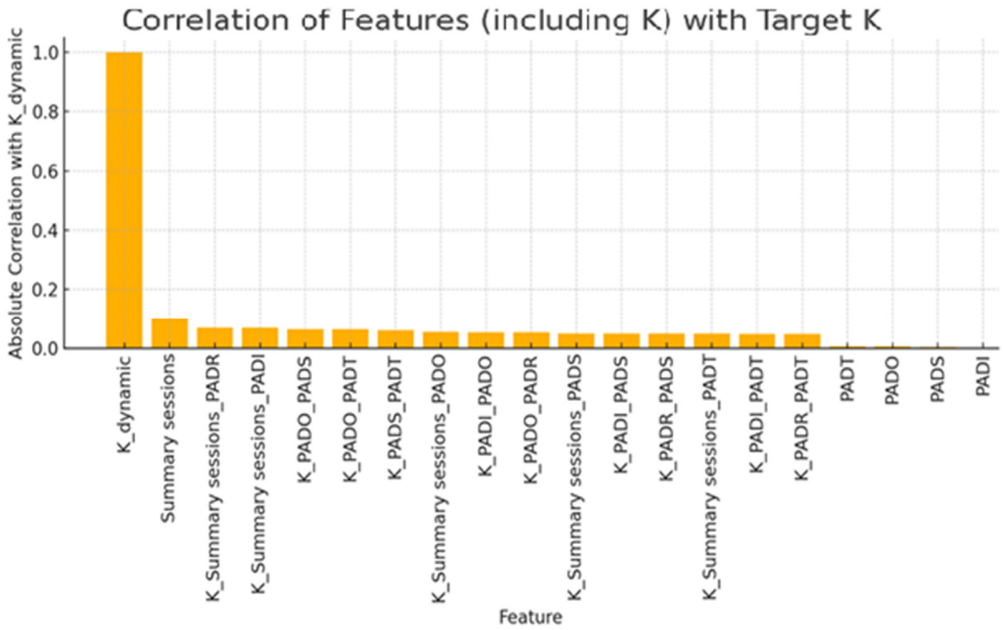

Fig. 2. Correlation of traits K_i with the coefficient K.

The analysis has demonstrated that it is most appropriate to use the dynamic attribute K for forecasting session losses, since it reflects the temporal dynamics of instability. Static attributes $K_{(i,j)}$ can usefully supplement the model, but are not its basis, as displayed in Figure 2. Afterwards, five regression models were trained on the processed PPPoE session data and their performance was evaluated using standard regression metrics: MSE and R^2. All models were evaluated using five-fold cross-validation to ensure robustness and generalizability.

Linear models, such as the Linear Regression and Ridge Regression, achieved similar MSE values (296.6), while Lasso Regression produced a higher MSE (773.3). Random Forest and SVR showed the largest MSE values (1699.8 and 1698.1, respectively), which is explained by the high variance and occasional spikes in the target variable *K*. Despite the similar MSE values between the Random Forest and SVR, their R^2 scores differed drastically. Random Forest achieved the highest R^2 (0.7445), meaning that it explains nearly 74% of the variance in the *K* coefficient. In contrast, SVR yielded a near-zero R^2 (−0.0005), indicating that its predictions do not improve over simply predicting the mean. This contrast demonstrates that MSE alone can be misleading without considering how well the model captures variance, as presented in Figures 3 and 4.

Linear Regression and Ridge Regression provided moderate R^2 values (≈ 0.4422), while Lasso was lower (≈ 0.3653). These results suggest that while simple linear models can partially capture relationships in the data, Random Forest is better suited to model the complex, nonlinear dependencies present in session dynamics. Kernel-based models, like SVR, appear to be unsuitable for this task, even when tuned. To further illustrate the model performance, a scatter plot was generated with the actual *K* values on the X-axis and the predicted values from the Random Forest model on the Y-axis. The points form a dense diagonal trend with visible spread, complying with the obtained R^2 score, as presented in Figure 5.

The distribution of points around the $y = x$ line indicates that the model captures the overall pattern well, though deviations occur due to the natural variability in the data. The results support the Random Forest as the most effective and robust model for this prediction task.

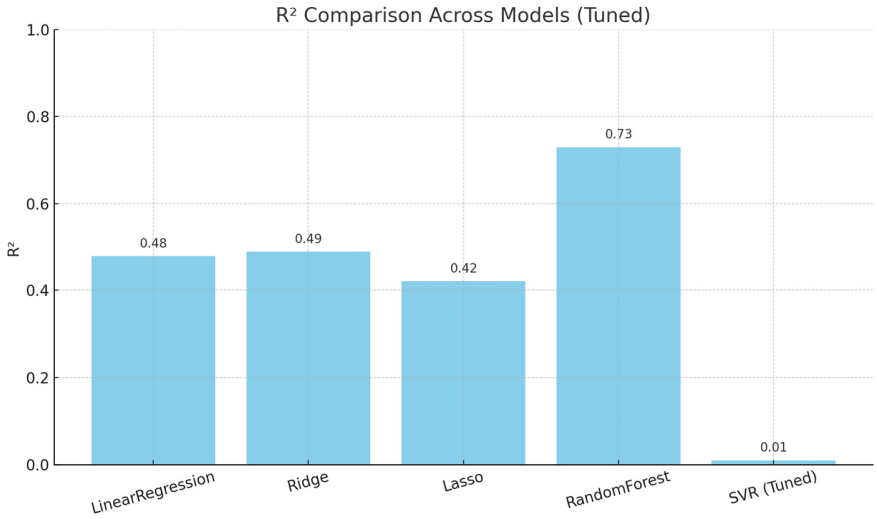

Fig. 3. Comparison of R^2 for each model.

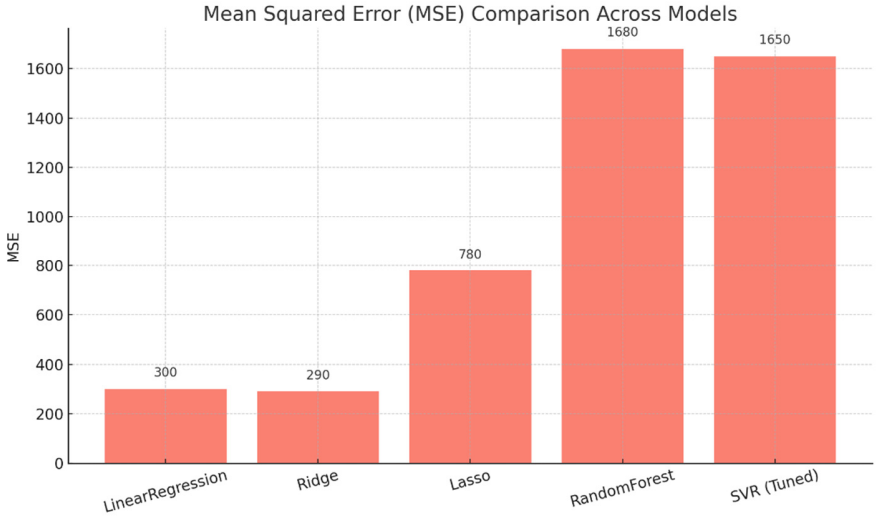

Fig. 4. Comparison of MSE across all models.

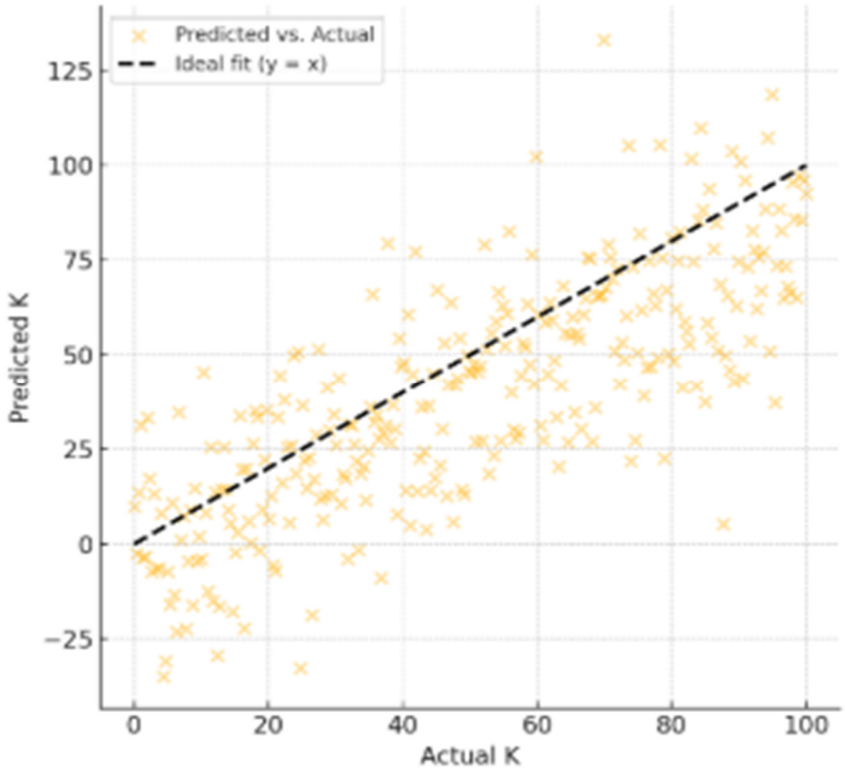

Fig. 5. Scatter plot of actual versus predicted *K* values using Random Forest.

```
gplearn location: C:\Program Files\Python\lib\site-packages\gplearn\__init__.py
    |   Population Average    |             Best Individual              |
---- ------------------------- ------------------------------------------ ----------
 Gen   Length          Fitness   Length          Fitness      OOB Fitness  Time Left
   0    26.36      2.29623e+89        7          24.5301              N/A     32.63s
   1    19.46      6.25252e+36        7          24.5301              N/A     14.51s
   2     7.69      2.88524e+45        7       0.00286581              N/A     15.90s
   3    10.89      2.21862e+42       15       0.00286581              N/A     15.85s
   4    12.40      4.24646e+25       15       0.00286581              N/A     13.53s
   5    16.80      1.29648e+30       19       0.00286482              N/A     13.46s
   6    12.82      7.40454e+38       19       0.00285805              N/A     14.55s
   7     6.03      1.80791e+36       17       0.00285687              N/A     11.67s
   8     5.12      7.05965e+51        7       0.00286581              N/A     10.57s
   9     5.03      3.78746e+40        5       0.00291074              N/A      9.11s
  10     5.04      2.66214e+42        5       0.00291074              N/A      8.50s
  11     5.08      6.35427e+56        5       0.00291074              N/A      8.08s
  12     5.11      8.82079e+36        5       0.00291074              N/A      6.39s
  13     5.24      8.41524e+56        5       0.00291074              N/A      5.95s
  14     5.25      1.84618e+27        5       0.00291074              N/A      4.33s
  15     5.18      3.14518e+31        5       0.00291074              N/A      3.59s
  16     5.09      4.90761e+35        5       0.00291074              N/A      2.93s
  17     5.21      4.96226e+32        5       0.00291074              N/A      1.76s
  18     5.10      1.82234e+38        5       0.00291074              N/A      1.04s
  19     5.21      2.77136e+26        5       0.00291074              N/A      0.00s

Best Program found:
div(sub(X2, X1), X0)

MSE on test=0.002962, R^2=0.999870
```

Fig. 6. Results of the symbolic regression experiment.

In addition to accuracy, the Random Forest model offers practical advantages for real-world deployment. Its inference time is low, and it operates solely on session-level counters (e.g., PADT, PADI), which are readily available in BRAS logs without requiring deep packet inspection. This makes the approach feasible for integration into real-time network monitoring systems.

Although deep learning models, such as LSTM networks, are widely used for time-dependent data, they require large volumes of labeled training data, significant tuning effort, and

substantial computational resources. Moreover, their lack of interpretability can be problematic in network operation environments. In contrast, the proposed method emphasizes the interpretability, efficiency, and minimal data requirements, offering a practical balance between the analytical rigor and deployment feasibility.

Symbolic regression methods allow algorithms to autonomously discover the mathematical relationships in data, offering a way to verify whether the K metric genuinely reflects the network issues without manually specifying formulas [30]. An experiment using the gplearn library (Figure 6) was conducted to search for expressions describing the dependency of K on Sessions, PADI, PADO, PADR, PADS, and PADT. The model successfully generated a formula closely matching (1), with $R^2 \approx 0.99987$ [21].

To validate the symbolic regression, the *SymbolicRegressor* (genetic programming) was used to recover this relationship. The model discovered an expression very close to the original formula, yielding an MSE of about 0.0030 and $R^2 \approx 0.9999$ on the test data. Figure 7 shows the quality of symbolic regression:

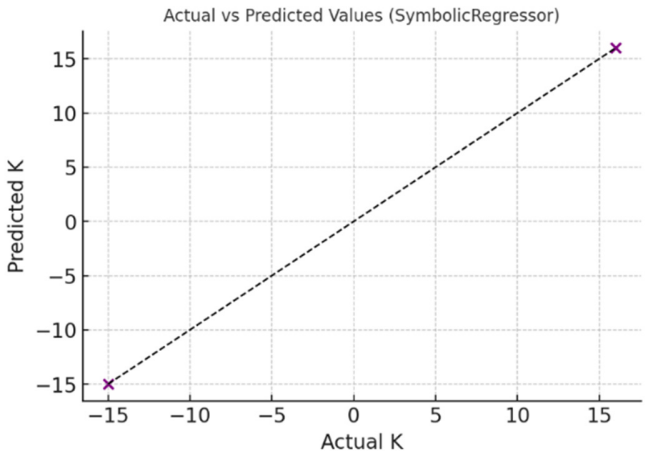

Fig. 7. Scatter plot of actual versus predicted *K* values of the *SymbolicRegressor* on synthetic data.

The points lie along the line $y=x$, indicating an outstanding model performance (R^2 close to 1, meaning that nearly 100% of the data variance is explained). The genetic algorithm for the symbolic regression gradually improved the solution with each generation. Figure 8 shows the change in the MSE of the best expression in the population across 20 generations. The error decreased rapidly in the initial few generations (from ~12 to <1), stabilizing around ~0.003 by generation 20, essentially reaching the noise level in the data.

MSE has rapidly decreased (orange curve, point mark generations), demonstrating the learning process and convergence toward an optimal formula. As a result, the model produced a formula matching (1), with $R^2 \approx 0.99987$. The symbolic regression algorithm (Genetic Programming) confirmed that a simple formula like this (or a closely related one) provides a minimal error when modeling the *K* coefficient. Thus, the hypothesis that *K* can be expressed through ΔPADT and sessions is validated. The metrics show a nearly perfect alignment, $R^2 \approx 0.99987$, indicating that the formula accurately reflects the underlying "law" of *K* dynamics. This approach aligns with previous work on the symbolic representations in logic-based signal processing systems [32, 33]. This method (symbolic regression) can also be applied in other domains requiring the automatic discovery of mathematical relationships in network statistics or similar datasets.

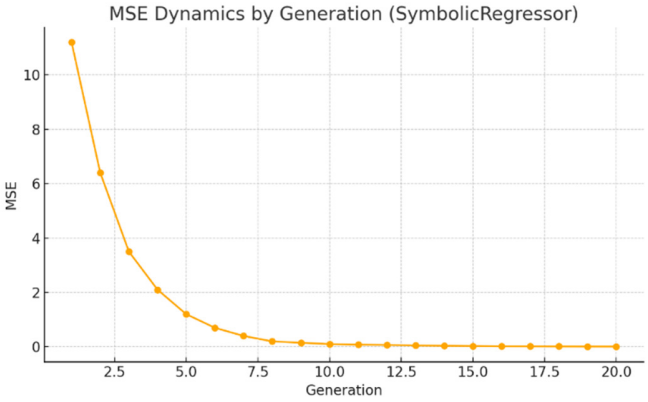

Fig. 8. MSE evolution of the best solution across generations for SymbolicRegressor.

IV. CONCLUSION

This study is the first to propose and empirically verify a method of indirect network quality monitoring based on analyzing the dynamics of PPPoE Active Discovery Termination (PADT) packets and the number of active Point-to-Point Protocol over Ethernet (PPPoE) sessions, enabling the calculation of the failure indicator *K*. Statistical analysis (Analysis of Variance (ANOVA)-test, correlation) identified a limited set of significant features, validating the effectiveness of feature pre-selection before applying machine learning algorithms.

Among the five regression models tested, the Random Forest model demonstrated the best performance for predicting the K coefficient, achieving a Coefficient of Determination (R^2) of approximately 0.7445 and outperforming linear models ($R^2 \approx 0.44$) and Support Vector Regression (SVR) ($R^2 \approx -0.0005$), even after parameter tuning. This confirms that ensemble-based models are better suited for capturing the nonlinear patterns present in network session behavior. The Mean Squared Error (MSE) of Random Forest was approximately 1699.8, which is acceptable given the high variance and occasional spikes in the target variable.

The application of symbolic regression further confirmed the mathematical validity of the original formula for *K*, supporting the hypothesis that session-level counters can effectively reflect link instability. The results demonstrate that integrating machine learning techniques into network monitoring pipelines can significantly improve the detection of anomalies and automate the failure prediction processes.

The practical value of the proposed approach lies in its ability to detect quality degradation early, using only aggregated session-level statistics without inspecting the traffic content. Future work may focus on log-transforming the target variable to reduce volatility, expanding the feature set with temporal dependencies, and deploying the model in real-time systems for continuous monitoring of the telecommunication infrastructure.

Although deep learning models, such as Long Short-Term Memory (LSTM) networks are frequently applied to temporal data, they were not included in this study due to several limitations: the need for large labeled datasets, higher computational overhead, and reduced interpretability. In contrast, the chosen regression models provide a transparent and resource-efficient solution, which is more compatible with operational environments where explainability and low-latency decision-making are critical.

Future work may explore log-transforming the target variable to improve stability, integrating time-series modeling techniques, and expanding the feature set with temporal dependencies, such as lags or cyclical patterns. Additionally, extending the proposed system into production-grade real-time monitoring platforms represents a promising direction for future research.

ACKNOWLEDGMENT

This research is funded by the JSC "Institute of Digital Engineering and Technology," Almaty, Kazakhstan.

REFERENCES

[1] H. Ren *et al.*, "Time-Series Anomaly Detection Service at Microsoft," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, Jul. 2019, pp. 3009–3017, https://doi.org/10.1145/3292500.3330680.

[2] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," *IEEE Access*, vol. 7, pp. 41525–41550, 2019, https://doi.org/10.1109/ACCESS.2019.2895334.

[3] L. Mamakos, K. Lidl, J. Evarts, D. Carrel, D. Simone, and R. Wheeler, "A Method for Transmitting PPP Over Ethernet (PPPoE)," RFC Editor, RFC2516, Feb. 1999, https://doi.org/10.17487/rfc2516.

[4] T. Babenko, H. Hnatiienko, and V. Vialkova, "Modeling of the Integrated Quality Assessment System of the Information Security Management System," in *CEUR Workshop Proceedings*, 2021, vol. 2845, pp. 75–84.

[5] J. Alkenani and K. Nassar, "Network Monitoring Measurements for Quality of Service: A Review," *Iraqi Journal for Electrical and Electronic Engineering*, vol. 18, no. 2, pp. 33–42, Dec. 2022, https://doi.org/10.37917/ijeee.18.2.5.

[6] Z. Ayan, B. Alimjan, M. Olga, Z. Timur, and Z. Toktalyk, "Quality of service management in telecommunication network using machine learning technique," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, Nov. 2023, Art. no. 1022, https://doi.org/10.11591/ijeecs.v32.i2.pp1022-1030.

[7] L. Chen and M. Zhao, "Machine Learning Techniques in QoS Management for PPPoE Networks," *Journal of Advanced Networking*, vol. 15, no. 2, pp. 45–56, Feb. 2021.

[8] P. Schummer, A. Del Rio, J. Serrano, D. Jimenez, G. Sánchez, and Á. Llorente, "Machine Learning-Based Network Anomaly Detection: Design, Implementation, and Evaluation," *AI*, vol. 5, no. 4, pp. 2967–2983, Dec. 2024, https://doi.org/10.3390/ai5040143.

[9] M. Yakubova, O. Manankova, A. Mukasheva, A. Baikenov, and T. Serikov, "The Development of a Secure Internet Protocol (IP) Network Based on Asterisk Private Branch Exchange (PBX)," *Applied Sciences*, vol. 13, no. 19, Sep. 2023, Art. no. 10712, https://doi.org/10.3390/app131910712.

[10] A. Zhunussov, A. S. Baikenov, and D. Ilieva, "Monitoring the quality of services provided in a telecommunication network by analyzing the statistics of PPPoE packets," in *2020 7th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE)*, Ruse, Bulgaria, Nov. 2020, pp. 1–4, https://doi.org/10.1109/EEAE49144.2020.9279089.

[11] Y. Gujarathi and Y. Potekar, "Machine Learning in Network Traffic Analysis: Classification, Optimization, and Security," *International Journal for Research in Applied Science and Engineering Technology*, vol. 13, no. 4, pp. 455–459, Apr. 2025, https://doi.org/10.22214/ijraset.2025.68216.

[12] G. Sadikova, M. Amreev, O. Manankova, A. Mukasheva, and T. Serikov, "Analysis and Research of Tasks for Optimizing Flows in Multiservice Networks Based on the Principles of a Systems Approach," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 9, pp. 2811–2825, 2022.

[13] S. Velednitsky, "The Future of Network Monitoring: How AI and Machine Learning Are Changing the Game," *Security*, Feb. 2025. https://www.netflowlogic.com/the-future-of-network-monitoring-how-ai-and-machine-learning-are-changing-the-game/.

[14] J. Sanusi, S. Adeshina, A. M. Aibinu, O. Oshiga, R. Prasad, and A. Dayyabu, "Mobility Prediction Algorithms for Handover Management in Heterogeneous LiFi and RF Networks: An Ensemble Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18300–18306, Dec. 2024, https://doi.org/10.48084/etasr.8884.

[15] Z. Wang *et al.*, "Failure prediction using machine learning and time series in optical network," *Optics Express*, vol. 25, no. 16, Aug. 2017, Art. no. 18553, https://doi.org/10.1364/OE.25.018553.

[16] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: a comprehensive evaluation," *Proceedings of the VLDB Endowment*, vol. 15, no. 9, pp. 1779–1797, May 2022, https://doi.org/10.14778/3538598.3538602.

[17] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long Short Term Memory Networks for Anomaly Detection in Time Series," in *the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, Arp 2015, pp. 89–94.

[18] L. Bounia and I. Setitra, "Computing Improved Explanations for Random Forests: k-Majoritary Reasons:," in *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, Porto, Portugal, 2025, pp. 188–198, https://doi.org/10.5220/0013143100003890.

[19] M. Z. Yakubova, O. A. Manankova, K. A. Tashev, and G. S. Sadikova, "Methodology of the Determining for Pearson's Criterion based on Researching the Value of Delays in the Transmitting of Information over a Multiservice Network," in *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, Tashkent, Uzbekistan, Nov. 2020, pp. 1–5, https://doi.org/10.1109/ICISCT50599.2020.9351419.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[21] P. Wyrwiński and K. Krawiec, "Learning Semantics-aware Search Operators for Genetic Programming." arXiv, 2025, https://doi.org/10.48550/ARXIV.2502.04568.

[22] J. Brownlee, "How to Perform Feature Selection With Numerical Input Data," *MachineLearningMastery.com*, Jun. 04, 2020. https://www.machinelearningmastery.com/feature-selection-with-numerical-input-data/.

[23] V. N. Vapnik, "Complete Statistical Theory of Learning," *Automation and Remote Control*, vol. 80, no. 11, pp. 1949–1975, Nov. 2019, https://doi.org/10.1134/S000511791911002X.

[24] M. Ali, I. Ullah, W. Noor, A. Sajid, A. Basit, and J. Baber, "Predicting the Session of an P2P IPTV User through Support Vector Regression

(SVR)," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 6021–6026, Aug. 2020, https://doi.org/10.48084/etasr.3635.

[25] Cisco Systems, *PPPoE Subscriber Management Configuration Guide*. San Jose, California, United States: Cisco, 2021.

[26] K. Venkatachalam, P. Prabhu, B. S. Balaji, M. Abouhawwash, and R. Rajadevi, "Recursive Feature Elimination with Ridge Regression (L2) Machine Learning Hybrid Feature Selection Algorithm for Diabetic Prediction using Random Forest Classifer." *In Review*, Jul. 23, 2021, https://doi.org/10.21203/rs.3.rs-742641/v1.

[27] A. C. Cardall, R. C. Hales, K. B. Tanner, G. P. Williams, and K. N. Markert, "LASSO (L1) Regularization for Development of Sparse Remote-Sensing Models with Applications in Optically Complex Waters Using GEE Tools," *Remote Sensing*, vol. 15, no. 6, Mar. 2023, Art. no. 1670, https://doi.org/10.3390/rs15061670.

[28] S. Obata, C. J. Cieszewski, R. C. Lowe, and P. Bettinger, "Random Forest Regression Model for Estimation of the Growing Stock Volumes in Georgia, USA, Using Dense Landsat Time Series and FIA Dataset," *Remote Sensing*, vol. 13, no. 2, Jan. 2021, Art. no. 218, https://doi.org/10.3390/rs13020218.

[29] Y. Kovalova, T. Babenko, O. Oksiiuk, and L. Myrutenko, "Optimization of Lifetime in Wireless Monitoring Networks," *International Journal of Computing*, vol. 19, no. 2, pp. 267–272, Jun. 2020, https://doi.org/10.47839/ijc.19.2.1770.

[30] M. Quade, M. Abel, K. Shafi, R. K. Niven, and B. R. Noack, "Prediction of dynamical systems by symbolic regression," *Physical Review E*, vol. 94, no. 1, Jul. 2016, Art. no. 012214, https://doi.org/10.1103/PhysRevE.94.012214.

[31] J. Gao, "R-Squared (R^2) – How Much Variation Is Explained?," *Research Methods in Medicine & Health Sciences*, vol. 5, no. 4, pp. 104–109, Sep. 2024, https://doi.org/10.1177/26320843231186398.

[32] E. S. Vitulyova, D. K. Matrassulova, and I. E. Suleimenov, "Construction of Generalized Rademacher Functions in Terms of Ternary Logic: Solving the Problem of Visibility of Using Galois Fields for Digital Signal Processing," *International Journal of Electronics and Telecommunications*, vol. 68, no. 2, pp. 237–244, Dec. 2021, https://doi.org/10.24425/ijet.2022.139873.

[33] I. E. Suleimenov, Y. S. Vitulyova, and D. K. Matrassulova, "Features of digital signal processing algorithms using Galois fields GF(2n+1)," *PLOS ONE*, vol. 18, no. 10, Oct. 2023, Art. no. e0293294, https://doi.org/10.1371/journal.pone.0293294.