ISSN: 2229-6948(ONLINE)
DOI: 10.21917/ijct.2026.0570

WEIBULL DISTRIBUTION-BASED ROBUST DEFENSE MECHANISM AGAINST POISONING ATTACKS IN FEDERATED LEARNING INTRUSION DETECTION SYSTEMS

N. Sudhir Reddy[1] and Shaik Jakeer Hussain[2]

[1]*Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Malla Reddy College of Engineering India*

[2]*Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Institute of Aeronautical Engineering, Hyderabad, India*

Abstract

Federated Learning (FL) has emerged as a promising paradigm for Network Intrusion Detection Systems (NIDS) by enabling distributed model training without sharing raw data. However, FL remains highly vulnerable to poisoning attacks, where malicious clients manipulate local updates to degrade global model performance. This study introduces WeiDetect, a Weibull distribution-based defense framework designed to identify and mitigate poisoning behaviors in FL-based NIDS environments. The proposed method models the statistical distribution of client-level gradient deviations using the Weibull probability distribution. By characterizing update irregularities through shape and scale parameters, WeiDetect distinguishes benign clients from adversarial participants without requiring labeled attack data. The framework integrates a robust aggregation mechanism that dynamically assigns weights to client updates based on their likelihood of conformity to learned Weibull behavior patterns. Experimental evaluation conducted on benchmark intrusion detection datasets demonstrates that WeiDetect consistently improves robustness under non-IID settings and various poisoning intensities. The model achieves higher detection accuracy and improved F1-score compared to conventional aggregation strategies such as FedAvg, trimmed mean, and median-based FL defenses. Additionally, WeiDetect maintains stable convergence behavior even under high adversarial participation rates, indicating strong resilience to model manipulation attacks. Experimental evaluations on benchmark intrusion datasets demonstrate that WeiDetect achieves 93.0% accuracy, 93.0% precision, 92.8% recall, and 92.9% F1-score under 30% poisoning conditions, significantly outperforming FedAvg, Trimmed Mean, and clustering-based defenses. Additionally, WeiDetect maintains a robustness index above 0.93, indicating strong stability under adversarial participation and non-IID data distributions.

Keywords:

Federated Learning, Poisoning Attacks, Network Intrusion Detection System, Weibull Distribution, Robust Aggregation

1. INTRODUCTION

Federated Learning (FL) has gained significant attention as a distributed machine learning paradigm that enables collaborative model training across multiple clients without exposing raw data. In the context of cybersecurity, FL-based Network Intrusion Detection Systems (NIDS) have been widely explored due to their ability to leverage decentralized network traffic data while preserving privacy. Early studies highlight that FL enhances scalability and privacy-preserving analytics in distributed environments [1–3]. However, despite these advantages, FL introduces new security vulnerabilities that are not present in centralized learning frameworks.

One of the primary challenges in FL-based NIDS is the presence of adversarial participants who can manipulate local model updates. Poisoning attacks, in particular, represent a severe threat where malicious clients intentionally inject corrupted gradients or manipulated training samples to distort the global model. These attacks are difficult to detect because FL inherently assumes that all participating clients are trustworthy. Moreover, the server does not have direct access to raw data, making traditional anomaly detection techniques less effective [4–5].

Another major challenge arises from the non-IID (non-independent and identically distributed) nature of network traffic data across clients. In real-world deployments, each client observes different traffic patterns depending on its network environment. This heterogeneity complicates the distinction between benign statistical variation and malicious update behavior. Consequently, robust detection of poisoning attacks becomes significantly more complex in distributed intrusion detection scenarios [4–5].

The central problem addressed in this study is the lack of statistically grounded defense mechanisms capable of distinguishing malicious updates from legitimate but heterogeneous client contributions in FL-based NIDS. Existing aggregation methods such as FedAvg are highly sensitive to adversarial manipulation, while robust alternatives like median or trimmed mean approaches often degrade performance under high data heterogeneity. Therefore, there is a need for a probabilistic and adaptive defense mechanism that can dynamically model update behavior without relying on labeled attack data [6].

To address this gap, the primary objective of this research is to develop a robust FL defense framework that can effectively identify poisoning clients while maintaining model accuracy under non-IID conditions. The proposed method, termed WeiDetect, leverages the Weibull distribution to model the statistical behavior of client update deviations. The Weibull distribution is particularly suitable due to its flexibility in modeling skewed and heavy-tailed data distributions, which are commonly observed in adversarial FL environments.

The novelty of WeiDetect lies in its distribution-aware anomaly characterization mechanism. Instead of relying on fixed thresholds or distance-based heuristics, the method learns probabilistic boundaries of normal client behavior using Weibull parameters. This enables adaptive identification of outlier updates even in highly dynamic and heterogeneous network conditions. Additionally, WeiDetect integrates a weighted aggregation strategy that continuously adjusts client influence based on their probabilistic conformity score.

The contributions of this work can be summarized as follows. First, we propose a Weibull distribution-based statistical modeling approach for detecting poisoning attacks in FL-based NIDS environments. Second, we design a robust aggregation framework that dynamically filters malicious updates without requiring labeled adversarial data. Third, we demonstrate through extensive evaluation that the proposed method improves both detection robustness and classification performance under varying attack intensities and non-IID data distributions. Finally, we provide empirical evidence that probabilistic modeling offers a more stable and adaptive defense mechanism compared to conventional robust aggregation techniques.

2. RELATED WORK

Recent advancements in federated learning security for intrusion detection systems have focused primarily on robust aggregation and anomaly detection strategies. Early foundational works explored FedAvg-based learning for distributed NIDS, demonstrating its effectiveness in privacy-preserving environments but also revealing susceptibility to poisoning attacks [7]. These vulnerabilities motivated the development of robust aggregation techniques designed to reduce the influence of malicious updates.

One prominent direction involves distance-based filtering methods, where client updates are compared using Euclidean or cosine similarity metrics. Updates that deviate significantly from the global mean are either down-weighted or removed. Although these approaches provide a basic defense mechanism, they often struggle in non-IID settings where natural client variability can resemble adversarial behavior [8]. This limitation reduces their applicability in realistic network environments.

Another category of methods includes median-based and trimmed mean aggregation strategies. These techniques improve robustness by reducing sensitivity to extreme values in the update distribution. However, studies show that such deterministic methods degrade model performance when data heterogeneity is high, as they oversimplify the underlying statistical structure of client updates [9].

In addition, several research efforts have explored clustering-based defenses, where client updates are grouped based on similarity before aggregation. Malicious clusters are identified and excluded from training. While clustering improves resilience against coordinated attacks, its performance depends heavily on the choice of distance metrics and clustering parameters, making it unstable in dynamic FL environments [10].

More recent approaches have adopted machine learning-based anomaly detection techniques. These methods train separate models to identify malicious updates using features extracted from gradient statistics. Although effective in controlled environments, they require labeled attack data and additional computational overhead, which limits scalability in large-scale FL systems [11].

Game-theoretic frameworks have also been proposed to model interactions between honest and malicious clients. These methods attempt to reach equilibrium strategies that minimize adversarial impact. However, they often rely on strong assumptions about attacker behavior, which may not hold in real-world scenarios where attack strategies evolve dynamically [12].

Probabilistic modeling approaches have gained attention as a promising alternative. By modeling update distributions using statistical functions, these methods aim to capture uncertainty and variability in client behavior. Gaussian-based models have been widely explored, but they often fail to represent skewed or heavy-tailed distributions commonly observed in adversarial FL environments [13].

To address this limitation, researchers have started investigating non-Gaussian distributions for anomaly detection. Heavy-tailed distributions such as Student's t-distribution and extreme value theory-based models provide better flexibility in modeling outliers. However, these methods still lack adaptive weighting mechanisms that can dynamically adjust client contributions during training [14].

Finally, hybrid defense frameworks combining robust aggregation and anomaly detection have been proposed to improve resilience. These methods integrate statistical filtering with reinforcement learning or optimization-based strategies. While they show improved robustness, their complexity and training instability remain major concerns for practical deployment [15].

In contrast to these existing approaches, the proposed WeiDetect framework introduces a Weibull distribution-based modeling strategy that captures both skewness and variability in client updates. Unlike Gaussian assumptions, Weibull modeling provides greater flexibility in representing heterogeneous and adversarial distributions. Furthermore, the integration of probabilistic weighting enables continuous adaptation during training, reducing the reliance on rigid thresholds or labeled attack data. This positions WeiDetect as a more stable and statistically grounded defense mechanism for FL-based intrusion detection systems.

3. PROPOSED METHOD: WEIDETECT (WEIBULL DISTRIBUTION-BASED DEFENSE FOR FL-NIDS)

WeiDetect introduces a probabilistic defense mechanism for federated learning-based intrusion detection systems by modeling client update deviations using the Weibull distribution. The core idea is to treat each client update as a stochastic sample drawn from an unknown reliability distribution and then estimate its likelihood under a learned Weibull model. Clients with low likelihood values are interpreted as potential poisoning sources and are down-weighted during aggregation. Unlike deterministic filtering rules, this formulation allows the system to adapt naturally to heterogeneous (non-IID) network environments while maintaining sensitivity to adversarial perturbations.

3.1 LOCAL TRAINING AND UPDATE GENERATION

In WeiDetect, the federated learning process begins at the client side where each participant independently trains a local intrusion detection model using its private network traffic dataset. The training is fully decentralized, meaning no raw data is transmitted to the central server at any stage. Each client maintains a local copy of the global model parameters and refines them using gradient-based optimization over multiple epochs.

This stage forms the foundation of the entire federated pipeline because the quality and integrity of local updates directly influence global convergence and robustness.

Let the global model at communication round *t* be represented as $\mathbf{W}^{(t)}$. Each client *k* receives this model and performs local optimization using its dataset $\mathrm{D}_k = \{(\mathbf{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}$. The learning objective is defined as a empirical risk minimization problem:

$$\mathrm{L}_k(\mathbf{W}) = \frac{1}{n_k}\sum_{i=1}^{n_k} \ell\left(f(\mathbf{x}_i^{(k)}; \mathbf{W}), y_i^{(k)} \right) \tag{1}$$

where $f(\cdot)$ denotes the intrusion detection model (typically a deep neural network), and $\ell(\cdot)$ represents a loss function such as cross-entropy. The optimization is carried out using stochastic gradient descent (SGD), where the local update rule is expressed as:

$$\mathbf{W}_k^{(t,e+1)} = \mathbf{W}_k^{(t,e)} - \eta \nabla \mathrm{L}_k(\mathbf{W}_k^{(t,e)}) \tag{2}$$

where, *e* denotes the local epoch index, and η is the learning rate. After completing *E* local epochs, the client produces a final updated model:

$$\mathbf{W}_k^{(t+1)} = \mathbf{W}_k^{(t)} - \eta \sum_{e=0}^{E-1} \nabla \mathrm{L}_k(\mathbf{W}_k^{(t,e)}) \tag{3}$$

The corresponding model update sent to the server is computed as:

$$\Delta \mathbf{W}_k^{(t)} = \mathbf{W}_k^{(t+1)} - \mathbf{W}^{(t)} \tag{4}$$

This update vector encapsulates how each client's local data distribution influences the global model. In benign scenarios, these updates remain statistically consistent with other clients. However, in real-world adversarial environments, this assumption does not always hold.

3.2 ROLE OF ADVERSARIAL MANIPULATION IN LOCAL UPDATES

In poisoning attacks, malicious clients intentionally manipulate their training process or directly alter gradient updates before transmission. This can be represented as:

$$\Delta \mathbf{W}_k^{(t)} = \Delta \mathbf{W}_k^{(t)} + \boldsymbol{\delta}_k^{(t)} \tag{5}$$

where $\boldsymbol{\delta}_k^{(t)}$ is an adversarial perturbation designed to mislead the global aggregation process. This perturbation may take several forms, including label flipping, gradient scaling, or targeted backdoor injection. A more structured representation of poisoned local training can be expressed as:

$$\tilde{\mathrm{L}_k}(\mathbf{W}) = \mathrm{L}_k(\mathbf{W}) + \alpha \cdot \mathrm{L}_{attack}(\mathbf{W}) \tag{6}$$

where L_{attack} is an adversarial objective that forces the model toward incorrect decision boundaries, and αcontrols attack strength. The resulting gradient becomes:

$$\nabla \tilde{\mathrm{L}_k}(\mathbf{W}) = \nabla \mathrm{L}_k(\mathbf{W}) + \alpha \nabla \mathrm{L}_{attack}(\mathbf{W}) \tag{7}$$

This formulation highlights a critical issue: even if only a small fraction of clients are malicious, their gradients can disproportionately influence the global model due to the averaging nature of federated learning.

3.2.1 Statistical Interpretation of Local Updates:

From a statistical perspective, each client update can be interpreted as a random sample drawn from an underlying distribution:

$$\Delta \mathbf{W}_k^{(t)} \sim \mathrm{P}_{benign} \quad \text{or} \quad \mathrm{P}_{malicious} \tag{9}$$

The challenge in federated intrusion detection is that these two distributions overlap in practice, especially under non-IID conditions. This makes it difficult to distinguish legitimate variability from adversarial manipulation using simple thresholding techniques.

3.3 DEVIATION FEATURE EXTRACTION

After receiving model updates from all participating clients, WeiDetect transitions into a critical analytical stage where raw gradient information is transformed into statistically meaningful deviation features. This step is essential because direct comparison of high-dimensional weight vectors is neither stable nor computationally interpretable in federated environments. Instead, the system compresses update behavior into compact deviation descriptors that preserve both magnitude-based and directional anomalies.

Let the global model at round *t* be $\mathbf{W}^{(t)}$, and the update submitted by client *k* be $\Delta \mathbf{W}_k^{(t)}$. The first operation is the construction of a global reference update, defined as:

$$\Delta \bar{\mathbf{W}}^{(t)} = \frac{1}{K}\sum_{k=1}^{K} \Delta \mathbf{W}_k^{(t)} \tag{10}$$

This reference represents the central tendency of all participating clients in the current communication round. It acts as a statistical anchor against which individual deviations are measured. However, in adversarial settings, this mean itself may be slightly biased, so WeiDetect avoids relying on it as a final decision metric and instead uses it only as a relative baseline.

3.4 MAGNITUDE-BASED DEVIATION MODELING

The first feature extracted is the Euclidean deviation, which captures how far a client update lies from the collective behavior of the system. It is defined as:

$$D_k^{(t)} = \| \Delta \mathbf{W}_k^{(t)} - \Delta \bar{\mathbf{W}}^{(t)} \|_2 \tag{11}$$

This quantity reflects the energy difference between an individual client update and the global consensus direction. In benign scenarios, updates cluster closely around the mean, resulting in low deviation values. In contrast, poisoned clients often introduce exaggerated gradient shifts, leading to significantly higher norms.

However, raw deviation values are sensitive to scale variations across training rounds. To address this instability, WeiDetect applies normalization:

$$\hat{D}_k^{(t)} = \frac{D_k^{(t)} - \mu_D^{(t)}}{\sigma_D^{(t)} + \epsilon} \tag{12}$$

where $\mu_D^{(t)}$ and $\sigma_D^{(t)}$ denote the mean and standard deviation of deviations across all clients in round *t*, and ϵ is a numerical stabilizer. This transformation ensures that deviation scores

remain comparable across rounds, even when the overall training dynamics shift.

3.4.1 Directional Deviation Modeling:

Magnitude alone is insufficient for distinguishing sophisticated poisoning attacks, especially when adversaries scale their updates to mimic benign behavior. To address this limitation, WeiDetect introduces directional deviation, which evaluates alignment between each client update and the global update trend.

The cosine similarity between a client update and the global mean update is computed as:

$$\cos(\theta_k^{(t)}) = \frac{\Delta \mathbf{W}_k^{(t)} \cdot \Delta \bar{\mathbf{W}}^{(t)}}{\| \Delta \mathbf{W}_k^{(t)} \| \, \| \Delta \bar{\mathbf{W}}^{(t)} \|} \tag{13}$$

The corresponding directional divergence is defined as:

$$S_k^{(t)} = 1 - \cos(\theta_k^{(t)}) \tag{14}$$

A value close to zero indicates strong alignment with the global update direction, while higher values indicate conflicting or adversarial behavior. This measure is particularly effective against poisoning strategies that manipulate gradient direction rather than magnitude.

3.4.2 Feature Normalization and Joint Representation

Since magnitude and directional features operate on different numerical scales, WeiDetect constructs a unified feature representation:

$$\mathbf{z}_k^{(t)} = \left[\hat{D}_k^{(t)}, S_k^{(t)} \right] \tag{15}$$

To ensure statistical consistency before Weibull modeling, a second-level transformation is applied:

$$\tilde{\mathbf{z}}_k^{(t)} = \frac{\mathbf{z}_k^{(t)} - \boldsymbol{\mu}_z^{(t)}}{\boldsymbol{\sigma}_z^{(t)} + \hat{o}} \tag{16}$$

where normalization is performed component-wise. This step is important because Weibull-based estimation assumes that input observations follow a consistent scale distribution. Without this normalization, scale drift across FL rounds could distort parameter estimation.

3.4.3 Statistical Interpretation of Feature Space:

At this stage, each client is no longer represented by a high-dimensional gradient vector but by a compact statistical descriptor. The transformation can be interpreted as: $\Phi : \mathrm{R}^d \to \mathrm{R}^2$, where the original parameter space is projected into a deviation feature space. This projection is not arbitrary; it is designed to preserve two fundamental properties of adversarial behavior: amplification (captured by $\hat{D}_k^{(t)}$) and misalignment (captured by $S_k^{(t)}$).

Under benign conditions, the feature distribution tends to form a compact cluster with low variance. Under poisoning attacks, the distribution becomes heavy-tailed and skewed, which is precisely the statistical condition that motivates the later use of the Weibull distribution.

3.5 WEIBULL DISTRIBUTION MODELING OF CLIENT BEHAVIOR

This stage forms the statistical core of WeiDetect. After transforming each client update into a low-dimensional deviation feature vector, the system shifts from geometric interpretation to probabilistic modeling. The central assumption is that benign client behaviors follow a skewed, heavy-tailed distribution due to inherent non-IID data characteristics in federated learning environments. Instead of forcing a symmetric Gaussian assumption, WeiDetect models this variability using the Weibull distribution, which naturally captures asymmetry and extreme-value sensitivity.

Each client k at round t is represented by a deviation magnitude derived from its feature vector: $x_k^{(t)} = \| \tilde{\mathbf{z}}_k^{(t)} \|$. These scalar observations constitute the empirical dataset over which the Weibull distribution is fitted.

3.5.1 Weibull Probability Modeling of Client Updates:

The probability density function (PDF) of the Weibull distribution is defined as:

$$f(x; k, \lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} \exp\left(-\left(\frac{x}{\lambda} \right)^k \right), \quad x \geq 0 \tag{17}$$

where: $k>0$ is the shape parameter controlling distribution skewness and $\lambda>0$ is the scale parameter controlling spread. In WeiDetect, $x_k^{(t)}$ represents the deviation strength of a client update, and the goal is to estimate parameters (k, λ) that best describe the distribution of benign clients. The likelihood of observing all client deviations in a given round is:

$$\mathrm{L}(k, \lambda) = \prod_{k=1}^{K} f(x_k^{(t)}; k, \lambda) \tag{18}$$

Taking logarithm for numerical stability yields:

$$\log \mathrm{L}(k, \lambda) = K \log k - Kk \log \lambda + (k-1) \sum_{k=1}^{K} \log x_k^{(t)} - \sum_{k=1}^{K} \left(\frac{x_k^{(t)}}{\lambda} \right)^k \tag{19}$$

This formulation transforms client update behavior into a statistical inference problem rather than a heuristic filtering task.

3.5.2 Parameter Estimation via Maximum Likelihood:

Direct closed-form solutions for $k$and λdo not exist, so WeiDetect uses iterative optimization. The estimation process is defined by solving: $\frac{\partial \log \mathrm{L}}{\partial k} = 0, \quad \frac{\partial \log \mathrm{L}}{\partial \lambda} = 0$. Expanding the derivatives leads to nonlinear equations:

$$\frac{\partial}{\partial k} \log \mathrm{L} = \frac{K}{k} - K \log \lambda + \sum_{i=1}^{K} \log x_i - \sum_{i=1}^{K} \left(\frac{x_i}{\lambda} \right)^k \log \left(\frac{x_i}{\lambda} \right) = 0 \tag{20}$$

$$\frac{\partial}{\partial \lambda} \log \mathrm{L} = -\frac{Kk}{\lambda} + k \sum_{i=1}^{K} \frac{x_i^k}{\lambda^{k+1}} = 0 \tag{21}$$

These equations are solved iteratively using numerical methods such as Newton–Raphson or fixed-point iteration. In each communication round, parameters are updated dynamically, allowing the model to adapt to evolving attack patterns and shifting client distributions.

3.6 WEIBULL-BASED LIKELIHOOD INTERPRETATION OF CLIENTS

Once parameters are estimated, each client update is evaluated under the learned distribution. The probability density value for client k is: $p_k^{(t)} = f(x_k^{(t)}; k, \lambda)$. However, WeiDetect primarily relies on the survival function, which provides a more stable anomaly interpretation:

$$S_k^{(t)} = \exp\left(-\left(\frac{x_k^{(t)}}{\lambda}\right)^k\right) \tag{22}$$

This survival probability represents how likely a client is to belong to the benign distribution tail. High values indicate normal behavior, while low values suggest abnormal deviation. The anomaly score is therefore defined as: $A_k^{(t)} = 1 - S_k^{(t)}$. This formulation avoids hard thresholds and instead assigns continuous risk values to each client.

3.7 ADAPTIVE BEHAVIOR UNDER NON-IID CONDITIONS

A key advantage of Weibull modeling is its flexibility in handling heterogeneous distributions. In federated learning, client updates are rarely symmetric due to differences in local data distributions. Gaussian-based models assume fixed variance symmetry, which often leads to misclassification of benign but diverse clients.

In contrast, the Weibull shape parameter k dynamically adjusts skewness:

- $k<1$: heavy-tailed behavior (high heterogeneity or attack presence)
- $k\approx1$: exponential-like distribution (moderate variability)
- $k>1$: more concentrated benign behavior

This adaptability allows WeiDetect to distinguish between natural non-IID variability and malicious deviation more effectively than fixed-distribution models.

3.8 ROBUSTNESS AGAINST POISONING ATTACKS

Poisoning attacks typically manifest as outliers in the deviation space. Since Weibull distribution is sensitive to tail behavior, such outliers significantly reduce survival probability: $\lim_{x_k^{(t)} \to \infty} S_k^{(t)} \to 0$. This property ensures that even subtle adversarial manipulations eventually accumulate statistical penalty across rounds. Unlike deterministic filtering methods, this effect is gradual and cumulative, making it more resilient against stealth attacks. Additionally, because parameter estimation is performed every round, attackers cannot easily exploit fixed decision boundaries. Any change in attack strategy immediately affects the estimated distribution, which in turn recalibrates client scoring.

4. LIKELIHOOD-BASED ANOMALY SCORING

This stage converts the statistical representation learned through the Weibull model into an explicit trust evaluation mechanism for each client. While the previous step estimates the global behavior of benign updates, this phase focuses on measuring how far each individual client deviates from that learned distribution. The key idea is not to classify clients directly, but to quantify their probability of belonging to the benign population in a continuous manner. After estimating the Weibull parameters k (shape) and λ (scale), each client update is evaluated using its deviation magnitude $x_k^{(t)}$, previously derived from the normalized feature space. Instead of relying on hard thresholds, WeiDetect computes a probabilistic score that reflects how consistent each update is with expected benign behavior.

4.1 SURVIVAL FUNCTION AS A TRUST MEASURE

The core of this mechanism is the Weibull survival function, which measures the probability that a random variable exceeds a given value under the learned distribution. In WeiDetect, this is interpreted as the probability that a client update remains within benign behavioral limits:

$$P_k^{(t)} = \exp\left(-\left(\frac{x_k^{(t)}}{\lambda}\right)^k\right) \tag{23}$$

where, $P_k^{(t)}$ acts as a trust score. A higher value indicates strong alignment with benign behavior, while a lower value suggests abnormal deviation. This formulation is particularly useful because it naturally compresses extreme deviations into near-zero probabilities without requiring explicit cut-off thresholds. From a behavioral standpoint, benign clients tend to cluster around lower deviation values, resulting in survival probabilities close to 1. In contrast, poisoned or manipulated clients produce inflated deviation magnitudes, which exponentially reduce their survival probability. This exponential decay is critical because it ensures that even moderate adversarial deviations are penalized significantly.

4.2 TRANSFORMATION INTO ANOMALY SCORE

While the survival probability represents trust, WeiDetect also defines an explicit anomaly score to quantify risk: $A_k^{(t)} = 1 - P_k^{(t)}$. This transformation provides a direct interpretation of adversarial likelihood. A value near 0 indicates a highly reliable client, whereas values approaching 1 indicate strong suspicion of poisoning behavior. This dual representation is useful in practice because it allows both reinforcement (through trust scores) and penalization (through anomaly scores) within the same framework. Importantly, the transformation remains monotonic, preserving ranking consistency across clients.

4.2.1 Normalized Risk Distribution Across Clients

To ensure comparability across all participating clients in a given communication round, WeiDetect further normalizes anomaly scores:

$$R_k^{(t)} = \frac{A_k^{(t)}}{\sum_{j=1}^{K} A_j^{(t)} + \dot{o}} \quad (24)$$

This normalization step converts raw anomaly values into a probability-like distribution over all clients. The resulting $R_k^{(t)}$ values sum to 1 and can be interpreted as relative risk proportions. This formulation is particularly important in federated learning because it avoids absolute decision boundaries, which are often unstable under non-IID data conditions. Instead, clients are evaluated relative to the entire population in that round, making the system adaptive to dynamic shifts in update behavior.

4.2.2 Temporal Smoothing for Stability

In practical federated environments, client behavior may fluctuate across communication rounds due to stochastic training effects or temporary data shifts. To avoid overreacting to such variations, WeiDetect introduces temporal smoothing:

$$\tilde{A}_k^{(t)} = \alpha A_k^{(t)} + (1-\alpha) A_k^{(t-1)} \quad (25)$$

where $\alpha \in [0,1]$ controls the balance between current observation and historical behavior.

This mechanism ensures that a single anomalous round does not immediately penalize a client heavily. Instead, the system gradually adapts to persistent patterns of malicious behavior. This is particularly relevant in poisoning attacks that are intermittent or stealthy, where adversaries inject harmful updates sporadically to avoid detection. From a stability perspective, temporal smoothing reduces variance in anomaly scoring and prevents oscillations in client trust assignments, which can otherwise destabilize global model convergence.

4.2.3 Statistical Interpretation of Likelihood Scores

The likelihood-based scoring process can also be interpreted from a probabilistic inference perspective. Each client update is treated as an observation drawn from an unknown mixture of benign and malicious distributions:

$$x_k^{(t)} \sim \pi \mathrm{P}_{benign} + (1-\pi)\mathrm{P}_{malicious} \quad (26)$$

WeiDetect does not explicitly estimate this mixture but instead approximates the posterior likelihood of benign membership using the Weibull survival function:

$$P(\text{benign}|\ x_k^{(t)}) \approx \exp\left(-\left(\frac{x_k^{(t)}}{\lambda}\right)^k\right) \quad (27)$$

This implicit probabilistic interpretation is important because it avoids the need for labeled attack data, which is typically unavailable in federated intrusion detection settings.

4.3 ADAPTIVE WEIGHT ASSIGNMENT STRATEGY

The adaptive weight assignment stage translates probabilistic trust scores into actionable influence factors for global model aggregation. After estimating the benign likelihood of each client through Weibull-based survival analysis, WeiDetect does not directly accept or reject updates. Instead, it assigns continuous weights that determine how strongly each client contributes to the global model update. This design avoids abrupt exclusion, which is often unstable in federated learning, especially under non-IID data distributions.

At this stage, each client k is associated with a survival probability $P_k^{(t)}$, derived from the Weibull distribution. This probability reflects the statistical confidence that the client behaves similarly to benign participants. The central objective is to convert these probabilities into normalized aggregation weights while preserving relative trust ordering and ensuring numerical stability during training.

4.3.1 Probabilistic Weight Normalization

The most direct formulation of weight assignment is based on normalization of survival probabilities:

$$w_k^{(t)} = \frac{P_k^{(t)}}{\sum_{j=1}^{K} P_j^{(t)}} \quad (28)$$

This ensures that all client contributions sum to one:

$$\sum_{k=1}^{K} w_k^{(t)} = 1 \quad (29)$$

From a modeling perspective, this step converts raw probabilistic trust into a convex combination framework, where the global model becomes a weighted expectation of client updates. Clients with higher survival probabilities naturally dominate aggregation, while suspicious clients contribute minimally without being completely discarded.

This soft weighting mechanism is particularly important because in federated intrusion detection systems, even slightly deviating clients may still contain useful information due to heterogeneous traffic distributions.

4.3.2 Exponential Sensitivity Control

To enhance robustness against stealth poisoning attacks, WeiDetect introduces a nonlinear weighting transformation using exponential scaling:

$$w_k^{(t)} = \frac{\exp(-\beta A_k^{(t)})}{\sum_{j=1}^{K} \exp(-\beta A_j^{(t)})} \quad (30)$$

where, $A_k^{(t)}$ is the anomaly score and $\beta > 0$ is a sensitivity parameter. This formulation behaves similarly to a softmax function over negative anomaly scores. The parameter β controls the aggressiveness of filtering:

- Low β: smoother weighting, tolerant to variation
- High β: sharper suppression of anomalous clients

This flexibility is critical in federated learning because attack intensity is not constant across rounds. When poisoning is weak or sparse, aggressive filtering may harm convergence. When attacks are strong, stronger penalization is required. The exponential form also ensures numerical stability by preventing linear accumulation of anomalous influence, which is a limitation in simpler averaging schemes.

4.3.3 Temporal Trust Propagation

Federated environments exhibit temporal continuity, meaning client behavior is not independent across rounds. To capture this property, WeiDetect integrates historical trust into the weighting mechanism:

$$w_k^{(t)} = \gamma w_k^{(t-1)} + (1-\gamma)\frac{P_k^{(t)}}{\sum_{j=1}^{K} P_j^{(t)}} \quad (31)$$

where $\gamma \in [0,1]$ controls memory strength.

This formulation introduces inertia into trust evolution. Clients that consistently behave benignly accumulate higher long-term influence, while repeated deviations gradually reduce their contribution. Unlike static filtering approaches, this mechanism captures behavioral persistence, which is a strong indicator in identifying stealthy poisoning attacks that operate intermittently. From a system stability perspective, temporal smoothing also prevents sudden fluctuations in model aggregation weights, which could otherwise destabilize convergence in early training rounds.

4.3.4 Robustness-Oriented Weight Clipping

Although probabilistic weighting reduces the influence of malicious clients, extreme adversarial updates can still introduce instability if not properly bounded. To mitigate this, WeiDetect applies an implicit clipping mechanism during aggregation:

$$\Delta \mathbf{W}_k^{(t)} \leftarrow \frac{\Delta \mathbf{W}_k^{(t)}}{\max(1, \| \Delta \mathbf{W}_k^{(t)} \| / \tau)} \quad (32)$$

where τ is a predefined threshold controlling maximum allowable update magnitude. This ensures that no single client, regardless of its computed weight, can dominate the global update through extreme gradient scaling. The clipping operation acts as a secondary safety layer, complementing the probabilistic weighting mechanism.

4.3.5 Statistical Interpretation of Adaptive Weights

The weighting strategy can also be interpreted from a Bayesian perspective. Each client update is treated as an observation, and the weight represents an approximation of posterior belief in its reliability: $w_k^{(t)} \approx P(\text{benign} \mid x_k^{(t)})$. Thus, the global model update becomes an expectation over a latent reliability distribution: $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \mathrm{E}_{k \sim P(\text{benign})}[\Delta \mathbf{W}_k^{(t)}]$. This interpretation is important because it reframes federated aggregation as probabilistic inference rather than deterministic averaging. Instead of assuming all clients are equally trustworthy, WeiDetect explicitly models uncertainty in client behavior.

5. RESULTS AND DISCUSSION

The proposed WeiDetect framework is implemented using Python 3.10 with TensorFlow Federated for simulating distributed training. Experiments are conducted on a system equipped with Intel Core i7 processor, 32 GB RAM, and NVIDIA RTX 3060 GPU. The simulation environment models 50 federated clients under non-IID data partitioning. Communication rounds are fixed at 100 with local epochs set to 5 per client. The FL server executes aggregation after each round using Weibull-based weighting. The experiment uses a controlled poisoning ratio ranging from 0% to 40% to evaluate robustness under adversarial conditions.

Table.1. Simulation and Training Configuration

Parameter	Value
Framework	TensorFlow Federated
Clients	50
Communication Rounds	100
Local Epochs	5
Batch Size	32
Learning Rate	0.001
Attack Ratio	0%–40%

As shown in Table 1, the federated learning environment is configured under controlled heterogeneous and adversarial conditions.

5.1 PERFORMANCE METRICS

- **Accuracy** – Measures correct classification rate of intrusion detection.
- **Precision** – Evaluates proportion of correctly identified attacks among predicted attacks.
- **Recall** – Measures detection rate of actual attacks.
- **F1-Score** – Harmonic mean of precision and recall.
- **Robustness Index (RI)** – Measures performance degradation under poisoning attacks.

5.2 DATASET DESCRIPTION

Table.2. Dataset Characteristics

Dataset	Samples	Features	Classes	Description
NSL-KDD	125,973	41	5	Network intrusion benchmark dataset
UNSW-NB15	175,341	49	10	Modern attack traffic dataset
CICIDS 2017	283,074	78	15	Real-world cyber attack dataset

As shown in Table.2, multiple benchmark intrusion datasets are used to ensure generalization across diverse network attack patterns. The proposed method is compared with FedAvg, Trimmed Mean Aggregation, and Clustering-Based FL Defense. FedAvg represents standard aggregation without defense, Trimmed Mean removes extreme updates, and clustering-based defense groups similar updates to eliminate anomalies.

Table.3. Accuracy vs Poisoning Ratio

Poisoning (%)	FedAvg	Trimmed Mean	Clustering Defense	WeiDetect
0	94.2	94.0	94.1	95.0
5	90.1	91.5	92.3	95.1
10	86.4	88.2	90.0	94.8
15	82.0	85.0	87.6	94.3
20	78.3	82.1	85.2	93.9
25	74.0	79.4	83.0	93.5
30	70.5	76.0	80.2	93.0

As shown in Table.3, WeiDetect consistently maintains higher accuracy compared to baseline methods across all poisoning levels. At 0% poisoning, all methods perform similarly, with WeiDetect reaching 95.0%. However, as adversarial intensity increases, FedAvg degrades sharply to 70.5% at 30% poisoning, indicating strong vulnerability to malicious updates. Trimmed Mean shows moderate robustness, stabilizing at 76.0%, but still suffers from loss of useful gradient information. Clustering-based defense performs better, reaching 80.2%, due to its ability to separate anomalous updates. In contrast, WeiDetect maintains 93.0% accuracy even under severe attack conditions. The performance gap widens progressively with increasing poisoning ratio, showing a clear advantage of Weibull-based probabilistic filtering. The gradual decline in WeiDetect indicates controlled sensitivity rather than abrupt degradation, which is crucial in real-world intrusion detection systems. The improvement of nearly 13% over clustering-based methods at high attack levels confirms the effectiveness of statistical modeling in preserving model integrity under adversarial FL conditions.

Table.4. Precision vs Poisoning Ratio

Poisoning (%)	FedAvg	Trimmed Mean	Clustering Defense	WeiDetect
0	93.8	94.1	94.0	95.2
5	89.0	90.8	92.0	95.0
10	85.2	87.5	89.6	94.7
15	80.1	84.0	87.0	94.2
20	76.0	81.2	84.5	93.8
25	72.3	78.0	82.0	93.4
30	69.5	75.2	79.0	93.0

The Table.4 illustrates that WeiDetect significantly outperforms baseline methods in precision across all poisoning levels. FedAvg suffers from a substantial increase in false positives as attack ratio increases, dropping to 69.5% precision at 30% poisoning. This indicates that adversarial updates severely distort its decision boundary. Trimmed Mean improves stability slightly, achieving 75.2%, but still misclassifies benign updates due to rigid statistical trimming. Clustering-based defense performs better, maintaining 79.0%, but its dependency on similarity thresholds limits adaptability. WeiDetect achieves consistently high precision, maintaining 93.0% even at maximum attack intensity. This improvement is attributed to Weibull-based probabilistic weighting, which reduces influence of suspicious clients rather than removing them entirely. The smooth attenuation mechanism ensures that useful but slightly deviating updates are preserved, reducing false alarms. The consistent precision levels demonstrate that WeiDetect maintains a stable decision boundary even in adversarial conditions, making it highly suitable for intrusion detection environments where false positives can significantly impact operational efficiency.

Table.5. Recall vs Poisoning Ratio

Poisoning (%)	FedAvg	Trimmed Mean	Clustering Defense	WeiDetect
0	94.5	94.3	94.4	95.3
5	88.8	90.0	91.5	95.1
10	84.0	87.0	89.2	94.6
15	79.0	83.5	86.8	94.1
20	75.0	81.0	84.0	93.7
25	71.2	77.5	82.0	93.2
30	68.0	74.0	79.5	92.8

The Table.5 shows that WeiDetect maintains superior recall performance compared to baseline methods under increasing poisoning intensity. FedAvg experiences a sharp decline in recall, reaching 68.0% at 30% poisoning, indicating that many true intrusions are missed due to poisoned gradients. Trimmed Mean and clustering-based methods perform moderately better, achieving 74.0% and 79.5% respectively. However, both methods still struggle to retain sensitivity under heavy adversarial manipulation. WeiDetect consistently achieves above 92% recall, demonstrating strong capability in detecting intrusion patterns even under attack. The probabilistic Weibull framework ensures that malicious updates are down-weighted rather than completely discarded, preserving useful signal components. This leads to improved sensitivity in detecting true intrusion cases. The steady recall curve also indicates that WeiDetect avoids over-filtering, which is a common issue in rigid defense mechanisms. The results confirm that statistical modeling of update reliability enhances detection capability without sacrificing generalization across heterogeneous client environments.

Table.6. F1-Score vs Poisoning Ratio

Poisoning (%)	FedAvg	Trimmed Mean	Clustering Defense	WeiDetect
0	94.0	94.0	94.0	95.1
5	88.9	90.3	91.7	95.0
10	84.7	87.6	89.4	94.6
15	80.0	83.7	86.9	94.2
20	75.5	81.5	84.8	93.7
25	72.0	78.5	82.5	93.3
30	68.5	75.5	79.3	92.9

As observed in Table.6, WeiDetect achieves the highest F1-score across all poisoning levels, indicating balanced performance between precision and recall. FedAvg exhibits the steepest decline, falling to 68.5% at 30% poisoning, reflecting its inability to maintain both detection sensitivity and correctness. Trimmed Mean performs moderately, while clustering-based defense maintains better balance but still declines under high attack ratios. WeiDetect maintains a stable F1-score above 92%, demonstrating consistent trade-off management between false positives and false negatives. The Weibull-based weighting mechanism contributes significantly to this balance by assigning continuous reliability scores rather than binary classification decisions. This ensures that no abrupt loss of information occurs during aggregation. The stable F1-score curve confirms that WeiDetect is well-suited for real-world intrusion detection scenarios where both detection accuracy and reliability are equally critical.

Table.7. Robustness Index vs Poisoning Ratio

Poisoning (%)	FedAvg	Trimmed Mean	Clustering Defense	WeiDetect
0	0.95	0.95	0.95	0.96
5	0.88	0.90	0.92	0.96
10	0.80	0.85	0.89	0.95
15	0.72	0.82	0.87	0.95
20	0.65	0.78	0.84	0.94
25	0.60	0.74	0.82	0.94
30	0.55	0.70	0.80	0.93

The Table.7 demonstrates the robustness advantage of WeiDetect under adversarial conditions. FedAvg shows rapid degradation in robustness index, falling to 0.55 at high poisoning levels, indicating high susceptibility to adversarial manipulation. Trimmed Mean and clustering-based defenses perform better, but still exhibit gradual decline as attack intensity increases. In contrast, WeiDetect maintains a robustness index above 0.93 even at 30% poisoning. This indicates strong resilience against poisoning attacks in federated environments. The Weibull-based statistical modeling ensures that outlier updates are consistently assigned lower influence, preserving overall model stability. The smooth probabilistic attenuation prevents abrupt performance collapse, which is a common limitation in deterministic defense mechanisms. The results confirm that WeiDetect provides a more stable and adaptive defense boundary compared to conventional approaches.

5.3 DISCUSSION

Across all evaluated metrics, WeiDetect demonstrates consistent superiority over FedAvg, Trimmed Mean, and clustering-based defenses. The performance gap becomes more pronounced as poisoning intensity increases, indicating strong adversarial resilience. While FedAvg collapses under high attack conditions, and other baselines degrade gradually, WeiDetect maintains stable performance across accuracy, precision, recall, F1-score, and robustness index. The Weibull-based probabilistic modeling plays a central role in ensuring smooth degradation rather than abrupt failure. This is particularly important in federated intrusion detection systems where stability under attack is critical. The weighted aggregation strategy ensures that no single malicious client dominates the global model update. Additionally, the method preserves useful information from slightly deviating clients, which improves generalization under non-IID conditions. The results collectively validate that statistical distribution modeling provides a more reliable defense strategy than deterministic filtering or clustering methods. WeiDetect therefore establishes a strong balance between robustness and predictive performance in adversarial FL environments.

6. CONCLUSION

This study introduces WeiDetect, a Weibull distribution-based defense mechanism designed to enhance robustness in federated learning-based intrusion detection systems under poisoning attacks. The method models client update deviations using a probabilistic Weibull framework and assigns adaptive weights based on anomaly likelihood. Experimental evaluations across multiple benchmark datasets demonstrate that WeiDetect consistently outperforms FedAvg, Trimmed Mean, and clustering-based defenses. Under high poisoning intensity (30%), WeiDetect achieves 93.0% accuracy, 93.0% precision, 92.8% recall, and 92.9% F1-score, while maintaining a robustness index above 0.93. These results confirm its ability to sustain stable performance in adversarial and non-IID environments. Unlike deterministic filtering methods, WeiDetect provides smooth attenuation of suspicious updates rather than complete exclusion, preserving useful learning signals. The integration of Weibull distribution modeling enables adaptive and statistically grounded decision-making, making the framework resilient to evolving attack strategies. Overall, the proposed method offers a scalable and reliable defense solution for secure federated intrusion detection systems in distributed network environments.

REFERENCES

[1] Y.C. Lai, J.Y. Lin, Y.D. Lin, R.H. Hwang, P.C. Lin, H.K. Wu and C.K. Chen, "Two-Phase Defense against Poisoning Attacks on Federated Learning-based Intrusion Detection", *Computers and Security*, Vol. 129, pp 1-11, 2023.

[2] Z. Zhang, Y. Zhang, D. Guo, L. Yao and Z. Li, "SecFedNIDS: Robust Defense for Poisoning Attack against Federated Learning-based Network Intrusion Detection System", *Future Generation Computer Systems*, Vol. 134, pp. 154-169, 2022.

[3] T. Karthikeyan and K. Praghash, "Improved Authentication in Secured Multicast Wireless Sensor Network (MWSN) using Opposition Frog Leaping Algorithm to Resist Man-in-Middle Attack", *Wireless Personal Communications*, Vol. 123, No. 2, pp. 1715-1731, 2022.

[4] P. Takkalapally, N. Sharma, A. Jaggi, K. Hudani and K. Gupta, "Assessing the Applicability of Adversarial Machine Learning Approaches for Cybersecurity", *Proceedings of International Conference on Advances in Computation, Communication and Information Technology*, Vol. 1, pp. 431-436, 2024.

[5] A. Jaggi, P. Takkalapally, S.K. Rajaram, K. Hudani and N. Jiwani, "Investigating Fault-Tolerance Techniques for Protecting Cyber-Physical Systems", *Proceedings of International Conference on Advances in Computation, Communication and Information Technology*, Vol. 1, pp. 437-442, 2024.

[6] N. Sharma, A. Jaggi, P. Takkalapally and K. Hudani, "Analyzing Adaptive Intrusion Detection Systems for Improved Network Security", *Proceedings of International Conference on Advances in Computation, Communication and Information Technology*, Vol. 1, pp. 425-430, 2024.

[7] M. Ilyas, N.I. Kajla, M.G. Madden, G. Fan, C. Zhang and H.M.S. Badar, "PoisonShield-FL-NIDS: A Robust Defense Against Poisoning Attacks in Federated Learning Intrusion Detection", *IEEE Internet of Things Journal*, Vol. 13, No. 4, pp. 6105-6115, 2025.

[8] C. Zhang, S. Yang, L. Mao and H. Ning, "Anomaly Detection and Defense Techniques in Federated Learning: A Comprehensive Review", *Artificial Intelligence Review*, Vol. 57, No. 6, pp. 1-34, 2024.

[9] C. Lewis, V. Varadharajan and N. Noman, "Attacks against Federated Learning Defense Systems and their Mitigation", *Journal of Machine Learning Research*, Vol. 24, No. 30, pp. 1-50, 2023.

[10] G. Xia, J. Chen, C. Yu and J. Ma, "Poisoning Attacks in Federated Learning: A Survey", *IEEE Access*, Vol. 11, pp. 10708-10722, 2023.

[11] X. Li, Z. Qu, S. Zhao, B. Tang, Z. Lu and Y. Liu, "Lomar: A Local Defense against Poisoning Attack on Federated Learning", *IEEE Transactions on Dependable and Secure Computing*, Vol. 20, No. 1, pp. 437-450, 2021.

[12] M. Benmalek, M.A. Benrekia and Y. Challal, "Security of Federated Learning: Attacks, Defensive Mechanisms and Challenges", *International Information and Engineering Technology Association*, Vol. 36, No. 1, pp. 49-59, 2022.

[13] A. Khraisat, A. Alazab, M. Alazab, T. Jan, S. Singh and M.A. Uddin, "Securing Federated Learning: A Defense Strategy against Targeted Data Poisoning Attack", *Discover Internet of Things*, Vol. 5, No. 1, pp. 1-9, 2025.

[14] E. Nowroozi, I. Haider, R. Taheri and M. Conti, "Federated Learning Under Attack: Exposing Vulnerabilities through Data Poisoning Attacks in Computer Networks", *IEEE Transactions on Network and Service Management*, Vol. 22, No. 1, pp. 822-831, 2025.

[15] Y. Elgharieb, W. Alexan, M. El-Aasser and M.M. Ghantous, "SpyShield: A Spyfall Inspired Defense Mechanism against Poisoning Attacks in Federated Learning", *Scientific Reports*, Vol. 15, No. 1, pp. 1-19, 2025.