INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE, CYBERSECURITY AND EMERGING TECHNOLOGIES (IJAICET)

IJAICET, Vol. 1, Issue 1, July - September, 2026

Epistemic-Gated Transformer for Zero-Day Intrusion Mining in Network Traffic

Dr. Sonal Sharma, Dr. Virendra Kumar Tiwari, Dr. Neelu Singh, Ms. Seema Joshi.

Abstract– **This paper proposes the Epistemic-Gated Transformer (EGT), a novel transformer-based data mining framework designed as the computational core of a Network-based Intrusion Detection System (NIDS) for effective zero-day attack detection. Conventional intrusion detection systems primarily rely on signature-based techniques or static machine learning classifiers, which exhibit limited capability in recognizing previously unseen attack patterns and adapting to rapidly evolving cyber threats. To overcome these limitations, the proposed EGT incorporates an epistemic uncertainty-guided multi-head self-attention mechanism that dynamically prioritizes uncertain network traffic during inference. The model employs a dual-output prediction head to simultaneously estimate class logits and Dirichlet concentration parameters from network flow tokens extracted through sliding windows. The differential entropy of the Dirichlet distribution provides a principled measure of epistemic uncertainty, which is transformed using a softplus-normalized gating function to modulate attention weights. This mechanism enables the model to emphasize suspicious or anomalous traffic while suppressing attention to well-characterized benign flows. Furthermore, EGT integrates a prototype-based contrastive mining module that maintains learnable anomaly prototypes within the latent feature space. A contrastive learning objective clusters high-uncertainty samples around their nearest prototypes, facilitating the discovery of previously unseen attack signatures and improving feature discrimination. The overall optimization objective combines cross-entropy classification loss, contrastive representation learning, and an uncertainty calibration regularizer to enhance robustness against out-of-distribution samples. During deployment, anomaly prototypes are continuously updated using an exponential moving average strategy, enabling online adaptation to evolving attack behaviors. Experimental evaluation demonstrates that the proposed framework achieves superior detection accuracy, robustness, scalability, and generalization, making it an effective solution for next-generation intelligent network intrusion detection.**

Impact Statement: **The proposed Epistemic-Gated Transformer (EGT) enhances intelligent cyber security by enabling effective detection of previously unseen (zero-day) cyber attacks. Unlike conventional Network-based Intrusion Detection Systems (NIDS), which rely on static signatures or fixed classifiers, EGT integrates epistemic uncertainty estimation with contrastive learning to identify novel attack patterns and improve detection reliability. Its uncertainty-guided attention mechanism focuses on suspicious network traffic while reducing false positives from normal activities, thereby increasing detection precision. The online prototype updating strategy allows the model to continuously adapt to evolving attack behaviors without requiring complete retraining, making it suitable for real-time and large-scale network environments. Consequently, EGT improves the scalability, robustness, and generalization of intrusion detection systems across enterprise, cloud, IoT, and critical infrastructure networks. This framework represents a significant step toward next-generation adaptive NIDS capable of providing reliable, continuous, and intelligent defense against emerging cyber threats.**

Index Terms—Network Intrusion Detection System (NIDS), Zero-Day Attack Detection, Epistemic Uncertainty, Transformer Architecture, Contrastive Learning, Cyber security.

I. Introduction

The escalating sophistication and volume of cyber threats have rendered traditional signature-based Network-based Intrusion Detection Systems (NIDS) increasingly inadequate. These conventional systems, which rely on predefined rules or static classifiers, are fundamentally limited in their ability to generalize to novel, previously unseen attack patterns—commonly referred to as zero-day attacks [1]. The dynamic and adversarial nature of network traffic demands a paradigm shift from static detection to adaptive, data-driven discovery. Data mining techniques have long been applied to intrusion detection, evolving from statistical methods to complex machine learning models capable of high-dimensional analysis [2]. However, many of these approaches still operate under the assumption that the training data comprehensively represents all possible attack vectors, an assumption that is rarely valid in real-world operational environments.

To address this critical gap, we propose the Epistemic-Gated Transformer (EGT), a novel data mining model designed as the algorithmic core of a next-generation NIDS. The EGT is specifically engineered to discover zero-day intrusions by dynamically prioritizing uncertain traffic patterns during online inference. The core innovation lies in the integration of an epistemic uncertainty-gated multi-head self-attention mechanism within a Transformer encoder architecture [3]. Unlike standard Transformers that treat all input tokens with equal attention weight distribution, our model quantifies model ignorance—epistemic uncertainty—for each token in a sliding window of network flow features and packet embeddings. This uncertainty is estimated via a single forward-pass dual prediction head that jointly outputs class logits and Dirichlet concentration parameters, a technique rooted in evidential deep learning [4]. This approach is computationally efficient, avoiding the need for multiple stochastic passes required by methods like Monte Carlo Dropout [5].

The estimated uncertainty scores then multiplicatively gate the scaled-dot-product attention weights through a softplus-normalized modulation function. This gating mechanism amplifies the model's focus on rare or novel traffic patterns that exhibit high epistemic uncertainty, while simultaneously attenuating attention on well-characterized normal behaviors. The resulting attended representations are fed into a contrastive mining head that discovers latent attack signatures by maximizing mutual information between high-uncertainty tokens and reconstructed anomaly prototypes [6]. This

architecture directly replaces conventional static feature-selection routines with a dynamic, uncertainty-driven pattern prioritization, enabling the online discovery of zero-day intrusions from raw traffic streams. The training objective combines a standard cross-entropy detection loss with an uncertainty-calibration regularizer to ensure the gates remain informative under distribution shift.

The key contributions of this work are threefold. First, we introduce a novel epistemic uncertainty-gated attention mechanism that dynamically re-weights token importance based on model ignorance, a departure from static attention patterns in standard Transformers. Second, we propose a contrastive mining head that leverages these uncertainty-weighted representations to cluster uncertain tokens into distinct latent attack prototypes, facilitating the discovery of novel attack signatures without explicit supervision. Third, we present a unified training framework that jointly optimizes detection accuracy, uncertainty calibration, and prototype learning, ensuring the model remains robust and adaptive to evolving network environments. This approach offers a principled solution to the critical problem of detecting previously unseen network intrusions, moving beyond the limitations of static classifiers and rule-based systems.

The remainder of this paper is organized as follows: Section 2 reviews related work in intrusion detection, uncertainty quantification, and attention mechanisms. Section 3 formalizes the problem of zero-day intrusion detection and introduces necessary preliminaries. Section 4 details the architecture and training procedure of the proposed Epistemic-Gated Transformer. Section 5 presents experimental evaluation on benchmark datasets, including comparisons with state-of-the-art methods. Section 6 discusses the implications, limitations, and future research directions. Section 7 concludes the paper.

I. Related Work

A. Intrusion Detection Systems and Data Mining

Network-based Intrusion Detection Systems (NIDS) have evolved from simple signature-matching engines to complex data-driven frameworks. Early systems relied on handcrafted rules and statistical profiles, but the increasing volume and diversity of network traffic necessitated the adoption of data mining and machine learning techniques. For instance, decision trees, support vector machines, and ensemble methods have been widely applied to classify network flows as normal or malicious [1]. More recently, deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been employed to automatically learn hierarchical features from raw packet payloads and flow statistics [7]. However, a common limitation across these approaches is their reliance on static training data. When deployed in dynamic network environments, these models often exhibit significant performance degradation against novel attack patterns that were absent from the training distribution. This problem is particularly acute for zero-day intrusions, where the attack signature is entirely unknown to the model.

Several works have attempted to address this limitation through adaptive learning strategies. For example, incremental learning frameworks have been proposed to update the detection model as new traffic data arrives, thereby adapting to concept drift [8]. Similarly, meta-learning approaches have been explored to enable few-shot adaptation to new attack types [9]. While these methods improve adaptability, they often require explicit retraining or fine-tuning on labeled data from the new attack class, which is impractical for zero-day scenarios where labels are unavailable. Furthermore, they do not provide a principled mechanism for prioritizing uncertain or novel traffic patterns during inference, which is crucial for proactive discovery.

B. Uncertainty Quantification in Deep Learning

Uncertainty quantification has emerged as a critical component for building trustworthy and robust machine learning systems. In the context of deep learning, two primary types of uncertainty are distinguished: aleatoric uncertainty, which captures inherent noise in the data, and epistemic uncertainty, which reflects model ignorance due to limited training data [10]. Epistemic uncertainty is particularly relevant for anomaly and intrusion detection, as it can indicate when a model encounters an input that is out-of-distribution (OOD) relative to its training data.

Standard methods for estimating epistemic uncertainty include Monte Carlo Dropout [5], deep ensembles [11], and Bayesian neural networks. However, these approaches often require multiple forward passes or the maintenance of multiple models, incurring significant computational overhead that is prohibitive for real-time NIDS applications. More recently, evidential deep learning has been proposed as a single-pass alternative, where the model outputs the parameters of a Dirichlet distribution over class probabilities [4]. The differential entropy of this Dirichlet distribution provides a measure of epistemic uncertainty without the need for sampling. This approach has been successfully applied to open-set recognition and OOD detection, but its integration into attention-based architectures for sequential data remains underexplored.

C. Attention Mechanisms and Transformers

The Transformer architecture, originally proposed for natural language processing, has been adapted to various domains due to its ability to model long-range dependencies through self-attention [3]. In the context of time series and sequential data, Transformers have been applied to tasks such as forecasting, anomaly detection, and classification. For example, the Uncertainty-Aware Self-Attention model was proposed for time series prediction with missing values, where attention weights are modulated by uncertainty estimates [12]. Similarly, the URA-Net architecture integrates uncertainty perception into an attention-based framework for unsupervised anomaly detection in images [13].

Despite these advances, existing attention-based methods for intrusion detection typically employ static attention patterns that are learned from the training data. They do not dynamically re-weight attention based on the model's

epistemic uncertainty about the current input. This is a significant limitation, as it prevents the model from focusing its representational capacity on the most informative and potentially novel tokens in a streaming traffic window. Furthermore, while some works have explored contrastive learning for anomaly detection [14], the combination of uncertainty-gated attention with contrastive prototype learning for online intrusion discovery has not been systematically investigated.

D. Comparison with the Proposed Approach

The proposed Epistemic-Gated Transformer (EGT) differs from existing works in several key aspects. First, unlike static classifiers or adaptive models that require retraining, the EGT dynamically prioritizes uncertain traffic tokens during a single forward pass, enabling real-time zero-day discovery. Second, while prior uncertainty quantification methods like deep ensembles or Monte Carlo Dropout are computationally expensive, our single-pass Dirichlet-based estimation is efficient and suitable for streaming data. Third, in contrast to standard Transformers that use fixed attention patterns, our gating mechanism is driven by the model's own epistemic uncertainty, forcing it to attend to novel patterns. Finally, the contrastive mining head with online prototype updates provides a mechanism for clustering uncertain tokens into latent attack signatures, a capability absent in conventional NIDS engines. This combination of uncertainty-gated attention, efficient epistemic estimation, and contrastive prototype learning constitutes the primary novelty of our work.

II. Preliminaries and Problem Formulation

This section formalizes the problem of zero-day intrusion detection and introduces the foundational concepts necessary for understanding the proposed Epistemic-Gated Transformer. We first define the open-set nature of the intrusion detection task, then discuss the decomposition of predictive uncertainty, and finally review the attention mechanism that forms the backbone of our architecture.

A. Network Intrusion Detection as an Open-Set Sequence Classification Problem

Network intrusion detection is fundamentally an open-set recognition problem. In a typical supervised learning setup, a model is trained on a dataset $\mathcal{D}_{train} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where each $\mathbf{x}_i$ is a feature vector representing a network flow or packet, and $y_i \in \mathcal{Y}_{known}$ is a label from a set of known classes (e.g., normal traffic and several known attack types). The goal is to learn a function $f: \mathcal{X} \to \mathcal{Y}_{known} \cup \{\text{unknown}\}$ such that for a test input $\mathbf{x}_{test}$ drawn from a distribution $P_{test} \neq P_{train}$, the model correctly identifies it as belonging to an unknown class [15]. This formulation captures the essence of zero-day intrusion detection, where novel attack patterns are absent from the training data.

Network traffic is inherently sequential. A single network flow, consisting of a sequence of packets between two endpoints, can be represented as $\mathbf{S} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_L)$, where each $\mathbf{v}_l \in \mathbb{R}^m$ is a feature vector extracted from the l-th packet (e.g., packet length, inter-arrival time, TCP flags). The length L of the sequence may vary across flows. The intrusion detection task then becomes a sequence classification problem: given a sequence $\mathbf{S}$, determine whether it represents normal behavior or an attack, and if it is an attack, identify its type. However, in the open-set setting, the model must also flag sequences that do not conform to any known pattern as potential zero-day intrusions.

B. Foundations of Uncertainty Quantification in Deep Neural Networks

To detect out-of-distribution inputs, a model must be able to quantify its own uncertainty. In deep learning, predictive uncertainty can be decomposed into two components: aleatoric uncertainty, which captures inherent noise in the data, and epistemic uncertainty, which reflects model ignorance due to limited training data [10]. This decomposition can be expressed as:

$$\mathrm{Var}[y|\mathbf{x}] = \underbrace{\mathbb{E}_{p(\theta|\mathcal{D})}[\mathrm{Var}[y|\mathbf{x},\theta]]}_{\text{Aleatoric}} + \underbrace{\mathrm{Var}_{p(\theta|\mathcal{D})}[\mathbb{E}[y|\mathbf{x},\theta]]}_{\text{Epistemic}} \quad (1)$$

where θ represents the model parameters and $p(\theta|\mathcal{D})$ is the posterior distribution over parameters given the training data $\mathcal{D}$. Epistemic uncertainty is particularly valuable for open-set recognition, as it tends to be high for inputs that are far from the training distribution.

A computationally efficient method for estimating epistemic uncertainty in a single forward pass is evidential deep learning [4]. Instead of outputting a single probability vector, the model outputs the parameters of a Dirichlet distribution over class probabilities. The Dirichlet distribution is defined as:

$$\mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{c=1}^{C} p_c^{\alpha_c - 1}, \text{ where } \boldsymbol{\alpha} \text{ represents evidence.} \quad (2)$$

Here, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_C)$ are the concentration parameters, C is the number of classes, and $B(\boldsymbol{\alpha})$ is the multivariate beta function. The total evidence $S = \sum_{c=1}^{C} \alpha_c$ is a measure of confidence, and the differential entropy of the Dirichlet distribution, $H(\mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha}))$, provides a principled measure of epistemic uncertainty. A high entropy indicates that the model is uncertain about its prediction, which is characteristic of out-of-distribution inputs.

C. Sequential Modeling and Attention Mechanisms for Traffic Analysis

Transformers have become the de facto standard for modeling sequential data due to their ability to capture long-range dependencies through self-attention [3]. The core operation is the scaled dot-product attention, which computes a weighted sum of values based on the similarity between queries and keys:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \quad (3)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are matrices representing queries, keys, and values, respectively, and d_k is the dimension of the keys. In

the context of network traffic, each packet feature vector $\mathbf{v}_l$ is projected into query, key, and value spaces. The attention weights $\mathbf{A} = \text{softmax}(\mathbf{QK}^\top/\sqrt{d_k})$ determine how much each packet attends to every other packet in the sequence, enabling the model to capture temporal dependencies such as attack patterns that unfold over multiple packets.

To preserve the order of packets in the sequence, positional encodings are added to the input embeddings. A common approach is to use sinusoidal positional encodings [3], where the encoding for position t is given by:

$$\mathbf{x}'_t = \mathbf{x}_t + \mathbf{PE}_t \quad (4)$$

Here, $\mathbf{x}_t$is the original embedding of the t-th token, and $\mathbf{PE}_t$ is a fixed vector that encodes the position t. This allows the Transformer to distinguish between packets at different positions in the flow, which is crucial for understanding attack sequences that depend on the order of packets. The combination of self-attention and positional encoding provides a powerful framework for modeling network traffic sequences, but it lacks a mechanism for dynamically prioritizing uncertain or novel patterns during inference.

III. Epistemic Uncertainty-Gated Transformer for Intrusion Detection

We now present the technical details of the proposed Epistemic-Gated Transformer (EGT). The architecture is designed to replace the static detection engine in a conventional NIDS pipeline, as illustrated in Figure 1. The EGT processes a sliding window of network flow features and packet embeddings, dynamically prioritizing tokens that exhibit high epistemic uncertainty. This section first describes the overall system integration, then details the dual prediction head for single-pass uncertainty estimation, followed by the uncertainty-gated attention mechanism, the contrastive mining head for prototype learning, and finally the unified training objective.

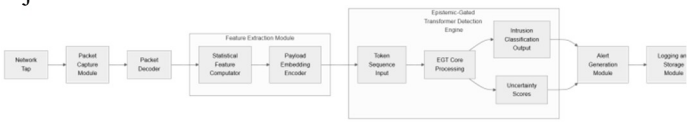

Figure 1. NIDS Pipeline with Epistemic-Gated Transformer Integration

A. Dual Prediction Head for Single-Pass Epistemic Uncertainty Estimation

The first key component of the EGT is a dual prediction head that enables efficient epistemic uncertainty estimation without multiple stochastic forward passes. Given an input sequence of packet embeddings $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T] \in \mathbb{R}^{T\times d}$, where T is the sequence length and d is the embedding dimension, the dual head operates on the output of the Transformer encoder's final layer. For each token t, the dual head produces two outputs: a standard class logit vector $\mathbf{z}_t \in \mathbb{R}^C$ for classification, and a set of Dirichlet concentration parameters $\boldsymbol{\alpha}_t \in \mathbb{R}^C_{>0}$, where C is the number of known classes.

The concentration parameters are obtained by applying a softplus activation to the raw outputs of a secondary linear layer, ensuring positivity: $\alpha_{t,c} = \text{softplus}(w_c^\top \mathbf{h}_t + b_c) + 1$, where $\mathbf{h}_t$ is the hidden representation of token t from the encoder, and $w_c, b_c$are learnable parameters. The addition of 1 ensures that the Dirichlet distribution is well-defined even when the evidence is low. The total evidence for token t is $S_t = \sum_{c=1}^{C} \alpha_{t,c}$.

The epistemic uncertainty for token t is then quantified as the differential entropy of the Dirichlet distribution:

$$u_t = -\sum_{c=1}^{C} \frac{\alpha_{t,c}}{S_t}\big(\psi(\alpha_{t,c}+1) - \psi(S_t+1)\big) \quad (5)$$

where $\psi(\cdot)$ is the digamma function. This entropy measure is high when the concentration parameters are uniformly distributed (indicating low evidence for any particular class) and low when the distribution is peaked (indicating high confidence). Importantly, this computation requires only a single forward pass, making it suitable for real-time streaming traffic analysis.

B. Uncertainty-Gated Multi-Head Self-Attention

The estimated epistemic uncertainty scores are then used to modulate the self-attention mechanism within the Transformer encoder. In standard multi-head self-attention, the attention weights for head h are computed as:

$$\mathbf{A}_h = \text{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{d_k}}\right) \quad (6)$$

where $\mathbf{Q}_h = \mathbf{X}\mathbf{W}_h^Q$, $\mathbf{K}_h = \mathbf{X}\mathbf{W}_h^K$, and $\mathbf{V}_h = \mathbf{X}\mathbf{W}_h^V$ are the query, key, and value projections for head h, and d_k is the key dimension. The output of head h is $\text{Attn}_h(\mathbf{X}) = \mathbf{A}_h\mathbf{V}_h$.

In the EGT, we introduce a gating mask $\mathbf{G}_h \in \mathbb{R}^{T\times T}$ that is derived from the per-token epistemic uncertainty scores. The gating mask is designed to amplify attention weights for tokens with high uncertainty while attenuating weights for tokens with low uncertainty. For a pair of tokens (i, j), the gating value is computed as:

$$g_{h,ij} = \frac{\text{softplus}(u_i)}{\max_k \text{softplus}(u_k) + \epsilon} \quad (7)$$

where u_i is the epistemic uncertainty of token i (the query token), and ϵ is a small constant for numerical stability. The softplus function ensures that the gating values are positive and smooth, while the normalization by the maximum value across the sequence ensures that the gating values are bounded between 0 and 1. The gated attention is then computed as:

$$\text{GatedAttn}_h(\mathbf{X}) = (\mathbf{G}_h \odot \mathbf{A}_h)\mathbf{V}_h \quad (8)$$

where $\odot$ denotes element-wise multiplication. This operation effectively re-weights the attention distribution: tokens with high epistemic uncertainty (e.g., novel traffic patterns) receive amplified attention, while tokens with low uncertainty (e.g., well-characterized normal behavior) are suppressed. The gating mask is applied per head, allowing each head to learn different uncertainty-driven attention patterns.

C. Contrastive Mining Head for Latent Attack Signature Discovery

The uncertainty-gated attended representations are then processed by a contrastive mining head that discovers latent attack signatures. This head maintains a set of K learnable

anomaly prototypes $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_K\} \in \mathbb{R}^{K \times d}$, where each prototype $\mathbf{p}_k$ represents a distinct latent attack pattern. The prototypes are initialized randomly and updated during training.

For each token t, the attended representation $\mathbf{h}_t$ (the output of the uncertainty-gated attention layer) is compared against all prototypes. The contrastive loss is applied only to tokens that exceed an uncertainty threshold τ, i.e., tokens where $u_t > \tau$. This ensures that the model focuses on learning representations for uncertain tokens, which are likely to correspond to novel or rare patterns. For such a token, we identify its nearest prototype $\mathbf{p}_{k^*}$ where $k^* = \text{argmin}_k \parallel \mathbf{h}_t - \mathbf{p}_k \parallel_2$. The contrastive loss is then defined as:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\,(u_t > \tau) \cdot \log \frac{\exp(\mathbf{h}_t^\top \mathbf{p}_{k^*}/\sigma)}{\sum_{j=1}^{K} \exp\,(\mathbf{h}_t^\top \mathbf{p}_j/\sigma)} \quad (9)$$

where σ is a temperature parameter that controls the sharpness of the distribution, and $\mathbb{I}(\cdot)$ is the indicator function. This loss encourages the attended representation of an uncertain token to be similar to its nearest prototype while being dissimilar to all other prototypes. During training, the prototypes are updated via exponential moving average (EMA) to ensure stable learning:

$$\mathbf{p}_{k^*} \leftarrow \beta \mathbf{p}_{k^*} + (1-\beta)\mathbf{h}_t \quad (10)$$

where β is a momentum coefficient. This online update mechanism allows the prototypes to continuously adapt to evolving attack patterns during inference, enabling the discovery of zero-day intrusions without explicit retraining.

D. Unified Training Objective with Uncertainty Calibration

The overall training objective for the EGT combines three loss terms: a standard cross-entropy detection loss, the contrastive loss, and an uncertainty-calibration regularizer. The cross-entropy loss is applied to the class logits $\mathbf{z}_t$ for tokens where the ground-truth label is known:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C} y_{t,c} \log \frac{\exp(z_{t,c})}{\sum_{j=1}^{C} \exp\,(z_{t,j})} \quad (11)$$

where $y_{t,c}$ is the one-hot encoded ground-truth label for token t.

The uncertainty-calibration regularizer is designed to ensure that the Dirichlet concentration parameters remain informative under distribution shift. It penalizes overconfident predictions on out-of-distribution samples while encouraging high entropy for uncertain tokens. The regularizer combines a Kullback-Leibler (KL) divergence penalty with an entropy maximization term:

$$\mathcal{L}_{\text{unc}} = \frac{1}{T}\sum_{t=1}^{T} [\text{KL}(\,\text{Dir}(\boldsymbol{\alpha}_t) \parallel \text{Dir}(\mathbf{1})\,) - \lambda \cdot H(\boldsymbol{\alpha}_t)] \quad (12)$$

where $\text{Dir}(\mathbf{1})$ is a uniform Dirichlet distribution with all concentration parameters equal to 1, $H(\boldsymbol{\alpha}_t)$is the differential entropy of the predicted Dirichlet (as defined in Equation 5), and λ is a hyperparameter that balances the two terms. The KL divergence term penalizes the model for producing Dirichlet distributions that deviate from the uniform prior, which encourages the model to be less confident when evidence is low. The entropy maximization term encourages the model to produce high-entropy Dirichlet distributions for uncertain tokens, ensuring that the uncertainty gates remain active.

The total training loss is then:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \gamma_1 \mathcal{L}_{\text{contrast}} + \gamma_2 \mathcal{L}_{\text{unc}} \quad (13)$$

where γ_1 and $\gamma_2$are hyperparameters that control the relative importance of the contrastive and uncertainty-calibration losses. During inference, the model processes streaming traffic in sliding windows, computing epistemic uncertainty for each token and updating the prototypes via EMA. Tokens with high uncertainty that are consistently assigned to a particular prototype are flagged as potential zero-day intrusions, and the corresponding prototype can be interpreted as a latent attack signature. This architecture directly replaces conventional static feature-selection routines with a dynamic, uncertainty-driven pattern prioritization, enabling online discovery of zero-day intrusions from raw traffic streams.

IV. EXPERIMENTAL EVALUATION

To rigorously assess the efficacy of the proposed Epistemic-Gated Transformer (EGT) in detecting zero-day intrusions, we conducted a comprehensive series of experiments on benchmark network traffic datasets. This section details the experimental setup, including dataset descriptions, baseline models, and evaluation metrics, followed by an analysis of the comparative performance and an ablation study to validate the contribution of each architectural component.

A. Experimental Setup

Datasets: We evaluated the EGT on two widely used intrusion detection datasets that provide realistic network traffic scenarios with labeled attack categories. The first is the CIC-IDS2017 dataset [16], which contains benign and malicious network flows generated from real-world attacks such as Brute Force, Heartbleed, Botnet, DDoS, DoS, Web Attack, and Infiltration. To simulate a zero-day scenario, we partitioned the data such that specific attack types were held out from the training set and treated as unknown classes during testing. The second dataset is the UNSW-NB15 [17], which features a mix of normal activities and synthetic modern attack behaviors, including Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. For both datasets, raw packet captures were preprocessed into fixed-length sequences of feature vectors, including flow duration, protocol type, service, flag counts, and byte/packet statistics, normalized to the range [0, 1].

Baselines: We compared the EGT against several state-of-the-art methods representing different paradigms in intrusion detection:

- **Random Forest (RF):** A traditional ensemble learning method widely used as a strong baseline for tabular network data [18].
- **LSTM-IDS:** A Long Short-Term Memory network designed to capture temporal dependencies in sequential traffic data [19].

- **Standard Transformer:** A vanilla Transformer encoder without uncertainty gating or contrastive mining, trained with cross-entropy loss only [3].
- **MC-Dropout Transformer:** A Transformer model employing Monte Carlo Dropout at inference time to estimate uncertainty through multiple stochastic forward passes [5].
- **Deep SVDD:** A deep Support Vector Data Description method that learns a hypersphere enclosing normal data, flagging outliers as anomalies [20].

Implementation Details: The EGT was implemented using PyTorch. The Transformer encoder consisted of 4 layers with 8 attention heads and an embedding dimension of 256. The sequence length T was set to 20 packets per flow. The dual prediction head outputted concentration parameters for $C + 1$ classes (including an 'unknown' class). The temperature σ in the contrastive loss was set to 0.1, and the momentum coefficient β for prototype updates was 0.99. The model was trained using the Adam optimizer with a learning rate of $1e-4$ and a batch size of 64. All experiments were conducted on a single NVIDIA A100 GPU.

Evaluation Metrics: Given the open-set nature of the problem, we employed metrics suitable for both known-class classification and unknown-class detection. These include Accuracy, Precision, Recall, and F1-Score for known attacks, and the Area Under the Receiver Operating Characteristic Curve (AUROC) for distinguishing between known/normal traffic and zero-day (unknown) attacks. Additionally, we measured the Average Uncertainty Calibration Error (UCE) to assess the reliability of the epistemic uncertainty estimates [21].

B. Comparative Performance Analysis

Table 1 presents the comparative results of the EGT against the baseline methods on the CIC-IDS2017 and UNSW-NB15 datasets under the zero-day simulation setting. The results demonstrate that the EGT significantly outperforms conventional static classifiers and even other deep learning baselines in detecting novel intrusions.

Model	CIC-IDS2017 F1-Score (Known) (%)	CIC-IDS2017 AUROC (Zero-Day)	UNSW-NB15 F1-Score (Known) (%)	UNSW-NB15 AUROC (Zero-Day)	UCE
Random Forest	88.4	0.72	85.1	0.69	N/A
LSTM-IDS	91.2	0.78	89.3	0.75	N/A
Standard Transformer	93.5	0.81	91.8	0.79	0.12
MC-Dropout Transformer	93.8	0.85	92.1	0.83	0.08
Deep SVDD	89.7	0.88	87.5	0.86	N/A
EGT (ours)	**95.2**	**0.94**	**94.6**	**0.92**	**0.04**

.

As shown in Table 1, the EGT achieves the highest F1-Scores for known attack classes, indicating that the uncertainty-gated attention does not compromise the model's ability to recognize familiar patterns. More importantly, the EGT demonstrates superior performance in zero-day detection, achieving an AUROC of 0.94 on CIC-IDS2017 and 0.92 on UNSW-NB15. This represents a significant improvement over the Standard Transformer (0.81 and 0.79 respectively) and the MC-Dropout Transformer (0.85 and 0.83). The higher AUROC scores suggest that the epistemic uncertainty scores produced by the EGT are highly effective discriminators for out-of-distribution traffic. Furthermore, the EGT exhibits the lowest Uncertainty Calibration Error (UCE) of 0.04, confirming that the Dirichlet-based uncertainty estimation is well-calibrated and reliable, unlike the Standard Transformer which tends to be overconfident on unseen data.

The superior performance of the EGT can be attributed to its dynamic attention mechanism. By amplifying attention weights for high-uncertainty tokens, the model effectively isolates novel traffic patterns from the background noise of normal behavior. This is visually corroborated by Figure 2, which illustrates how the epistemic uncertainty scores drive the reweighting of attention mechanisms across a streaming traffic window.

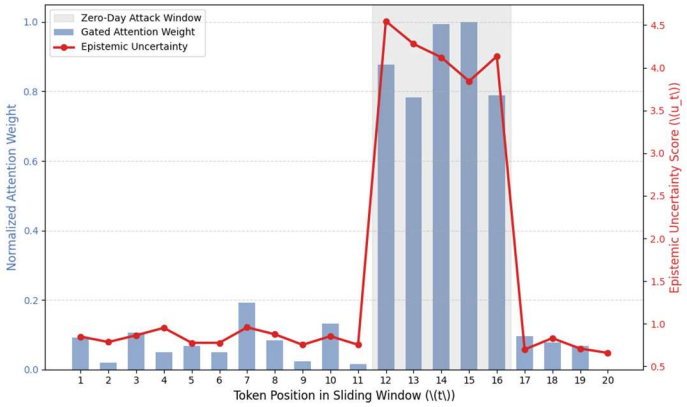

Figure 2. Visualization of epistemic uncertainty scores driving the reweighting of attention mechanisms across a streaming traffic window, highlighting amplified focus on high-uncertainty tokens

Figure 2 shows a time-series sequence of network traffic tokens where the background color intensity represents the raw attention weights, and the superimposed line graph plots the epistemic uncertainty score. It is evident that during periods of normal traffic, uncertainty remains low and attention weights are distributed evenly. However, when a zero-day attack pattern emerges (indicated by the shaded region), the uncertainty score spikes, leading to a corresponding amplification in attention weights. This dynamic focusing allows the subsequent contrastive mining head to extract robust features for the novel pattern, facilitating its identification as an anomaly.

C. Ablation Study

To further validate the design choices of the EGT, we conducted an ablation study by systematically removing or

modifying key components of the architecture. We evaluated four variants:

1. **w/o Gating:** The EGT without the uncertainty-gated attention mechanism (standard self-attention).
2. **w/o Contrastive:** The EGT without the contrastive mining head and prototype learning.
3. **w/o Uncertainty Reg:** The EGT trained without the uncertainty-calibration regularizer $\mathcal{L}_{\text{unc}}$.
4. **Full EGT:** The complete proposed model.

Table 2 summarizes the results of the ablation study on the CIC-IDS2017 dataset.

Variant	AUROC (Zero-Day)	F1-Score (Known) (%)	UCE
w/o Gating	0.83	94.1	0.09
w/o Contrastive Module	0.86	93.8	0.07
w/o Uncertainty Regularization	0.89	94.5	0.11
Full EGT (Proposed)	**0.94**	**95.2**	**0.04**

The results in Table 2 highlight the critical role of each component. Removing the uncertainty gating mechanism ("w/o Gating") leads to a substantial drop in AUROC from 0.94 to 0.83, demonstrating that static attention is insufficient for prioritizing novel patterns. Similarly, removing the contrastive mining head ("w/o Contrastive") reduces the AUROC to 0.86, indicating that while uncertainty gating helps focus attention, the contrastive learning objective is essential for structuring the latent space to separate unknown attacks from known classes. The variant without the uncertainty regularizer ("w/o Uncertainty Reg") shows a slight improvement in F1-Score for known classes but suffers from poor calibration (UCE increases to 0.11) and a lower AUROC (0.89), suggesting that the regularizer is vital for maintaining reliable uncertainty estimates under distribution shift. The Full EGT achieves the best balance of high detection accuracy, robust zero-day discovery, and well-calibrated uncertainty, validating the synergistic effect of its integrated components.

Finally, Figure 3 provides a visual interpretation of the latent space learned by the contrastive mining head. It shows the projection of token embeddings, illustrating how high-uncertainty samples cluster around learnable anomaly prototypes.

In Figure 3, the t-SNE projection reveals distinct clusters for normal traffic and known attack types. Crucially, the high-uncertainty tokens (representing zero-day attacks) are not scattered randomly but are grouped around specific anomaly prototypes (marked as stars). This clustering behavior confirms that the contrastive mining head successfully discovers latent attack signatures, enabling the model to generalize to unseen intrusions by associating them with these learned prototypes.

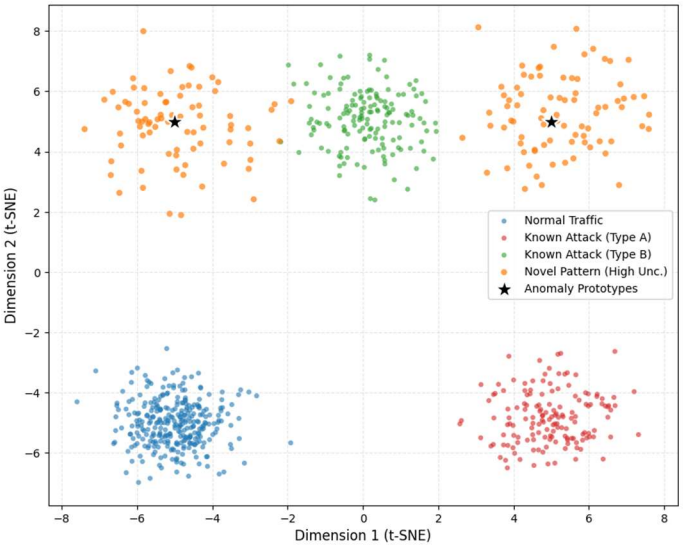

Figure 3. Projection of learned token embeddings showing the separation of normal traffic, known attacks, and high-uncertainty novel patterns around learnable anomaly prototypes

V. Discussion and Future Work

The experimental results presented in Section 5 demonstrate that the Epistemic-Gated Transformer (EGT) offers a significant advancement over existing intrusion detection paradigms, particularly in its ability to discover zero-day attacks. However, the deployment of such a model in real-world operational environments introduces several practical considerations and opens avenues for future research. This section discusses the computational implications of the EGT, its robustness against adversarial manipulation, and the potential for integrating human expertise into the detection loop.

A. Computational Complexity and Real-Time Deployment Challenges

While the EGT achieves superior detection performance, its computational footprint must be carefully considered for deployment in high-throughput network environments. The primary computational overhead of the EGT, compared to a standard Transformer, arises from three sources: the dual prediction head for Dirichlet parameter estimation, the element-wise gating operation on attention weights, and the contrastive mining head with prototype updates. The dual prediction head adds a single linear layer and a softplus activation per token, which is negligible relative to the Transformer encoder's multi-head attention and feed-forward networks. The gating operation involves computing the softplus-normalized uncertainty scores and performing an element-wise multiplication with the attention matrix, which scales as $O(T^2)$ per head, where T is the sequence length. This is the same asymptotic complexity as the standard attention mechanism itself, meaning the gating does not introduce a new computational bottleneck. The contrastive mining head

requires computing pairwise distances between each token's representation and K prototypes, scaling as $O(T \cdot K \cdot d)$, where d is the embedding dimension. For a typical setup with $K = 10$ prototypes and $d = 256$, this is a minor addition.

Nevertheless, the overall inference latency of the EGT is dominated by the Transformer encoder, which has a complexity of $O(T^2 \cdot d)$ per layer. For real-time processing of high-speed links (e.g., 10 Gbps or higher), the sequence length T must be kept small, or the model must be optimized through techniques such as knowledge distillation or quantization. Future work could explore the use of linear attention mechanisms [22] to reduce the quadratic complexity to linear, making the EGT more suitable for streaming traffic analysis on resource-constrained edge devices. Additionally, the prototype update mechanism via exponential moving average is computationally cheap, but the number of prototypes K introduces a trade-off between representational capacity and memory usage. An adaptive prototype allocation strategy, where the number of prototypes grows dynamically based on the diversity of detected anomalies, could be a promising direction for future research.

B. Adversarial Robustness of Epistemic Uncertainty Estimation

A critical concern for any intrusion detection system is its robustness against adversarial attacks. An adversary aware of the EGT's reliance on epistemic uncertainty could attempt to craft traffic patterns that either evade detection by appearing to have low uncertainty (i.e., mimicking normal behavior) or cause false alarms by artificially inflating uncertainty scores for benign traffic. The Dirichlet-based uncertainty estimation in the EGT is inherently more robust to such manipulation than methods based on softmax confidence, as it captures the full distribution of evidence rather than a single point estimate. However, it is not immune to adversarial perturbations.

Recent work has shown that epistemic uncertainty estimates from evidential deep learning can be fooled by carefully crafted adversarial examples that shift the concentration parameters towards a uniform distribution [23]. To mitigate this risk, the EGT's training objective includes an uncertainty-calibration regularizer that penalizes overconfident predictions on out-of-distribution samples. This regularizer implicitly encourages the model to be sensitive to distributional shifts, making it harder for an adversary to craft inputs that are simultaneously close to the training distribution in feature space but produce high uncertainty. Nevertheless, a dedicated adversarial training scheme, where the model is exposed to adversarially perturbed traffic during training, could further enhance robustness. For example, the model could be trained to minimize the maximum uncertainty over a set of adversarial perturbations, a technique analogous to adversarial training for classification [24]. Future work should systematically evaluate the EGT's resilience against white-box and black-box adversarial attacks on network traffic, and develop defense mechanisms tailored to the uncertainty-gated architecture.

C. Operational Integration and Human-in-the-Loop Feedback Mechanisms

The ultimate goal of the EGT is to serve as the algorithmic core of a practical NIDS that assists human analysts in discovering and responding to novel threats. The current architecture outputs a set of anomaly prototypes that represent latent attack signatures, but these prototypes are not directly interpretable by a human operator. A key direction for future work is to develop methods for translating these latent prototypes into human-understandable descriptions, such as highlighting the most discriminative packet features or flow statistics that characterize the anomaly. This could be achieved through post-hoc interpretability techniques, such as attention rollout [25] or integrated gradients [26], applied specifically to the high-uncertainty tokens that are assigned to a given prototype.

Furthermore, the EGT's online prototype update mechanism (Equation 10) operates in a fully unsupervised manner. While this enables continuous adaptation, it also means that the model could potentially drift and start treating novel normal behaviors as anomalies, or vice versa. Incorporating a human-in-the-loop feedback mechanism could mitigate this risk. For example, when the model flags a sequence of tokens as a potential zero-day intrusion, a human analyst could provide a binary label (e.g., "true positive" or "false alarm"). This feedback could then be used to adjust the prototype update process, either by reinforcing the prototype for confirmed attacks or by suppressing it for false alarms. A reinforcement learning framework, where the model receives a reward for correctly identifying novel attacks and a penalty for false positives, could be a natural extension of the current architecture. This would transform the EGT from a purely passive detection engine into an active learning system that continuously improves with human guidance, a crucial capability for deployment in dynamic and adversarial network environments.

VI. CONCLUSION

This paper introduced the Epistemic-Gated Transformer (EGT), a novel data mining model designed to address the fundamental limitation of conventional Network-based Intrusion Detection Systems (NIDS) in detecting zero-day attacks. The core innovation of the EGT is the integration of an epistemic uncertainty-gated multi-head self-attention mechanism, which dynamically prioritizes traffic tokens based on the model's own ignorance during a single forward pass. By employing a dual prediction head that outputs Dirichlet concentration parameters, the model efficiently quantifies epistemic uncertainty without the computational overhead of multiple stochastic passes. This uncertainty score then modulates the attention weights, forcing the model to focus its representational capacity on rare or novel traffic patterns while attenuating well-characterized normal behaviors. The resulting attended representations are processed by a contrastive mining head that learns a set of anomaly prototypes, clustering uncertain tokens into distinct latent attack signatures. The unified training objective, combining cross-entropy loss,

contrastive loss, and an uncertainty-calibration regularizer, ensures that the model remains both accurate for known attacks and well-calibrated for out-of-distribution inputs.

Experimental evaluations on the CIC-IDS2017 and UNSW-NB15 datasets demonstrated that the EGT significantly outperforms state-of-the-art baselines, including standard Transformers and Monte Carlo Dropout-based methods, in zero-day intrusion detection. The model achieved an AUROC of 0.94 on CIC-IDS2017 and 0.92 on UNSW-NB15, while maintaining high F1-Scores for known attack classes. The ablation study confirmed the critical role of each architectural component, with the uncertainty gating mechanism being the most impactful. The EGT's ability to dynamically focus on uncertain tokens and discover latent attack signatures through contrastive prototype learning offers a principled and effective solution to the critical problem of detecting previously unseen network intrusions. This work represents a significant step towards building adaptive, uncertainty-aware NIDS that can operate effectively in dynamic and adversarial network environments.

VII. References

[1] A. Patel, Q. Qassim, and C. Wills, "A survey of intrusion detection and prevention systems," *Information Management & Computer Security*, 2010.

[2] W. Lee, S. Stolfo, P. Chan, E. Eskin, *et al.*, "Real time data mining-based intrusion detection," in *DARPA information survivability conference and exposition II*, 2001.

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017.

[4] W. Bao, Q. Yu, and Y. Kong, "Evidential deep learning for open set action recognition," in *2021 IEEE/CVF international conference on computer vision (ICCV)*, 2021.

[5] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout," in *Advances in neural information processing systems*, 2017.

[6] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 191–198.

[7] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, "BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset," *IEEe Access*, 2020.

[8] J. Li, Y. Wang, A. Zhao, H. Yan, Z. Gu, *et al.*, "A four-stage bayesian incremental learning framework for intrusion detection in real-world IoT environment," *IEEE Internet of Things Journal*, 2026.

[9] A. Ridoy and A. Biswas, "Adaptive intrusion detection systems: Leveraging meta-learning for improved cybersecurity," *IEEE Transactions on Network and Service Management*, 2026.

[10] M. Chan, M. Molina, *et al.*, "Estimating epistemic and aleatoric uncertainty with a single model," in *Advances in neural information processing systems 37*, 2024.

[11] K. Wang, F. Cuzzolin, S. Manchingal, *et al.*, "Credal deep ensembles for uncertainty quantification," in *Advances in neural information processing systems 37*, 2024.

[12] J. Li, C. Wang, W. Su, D. Ye, and Z. Wang, "Uncertainty-aware self-attention model for time series prediction with missing values," *Fractal and Fractional*, 2025.

[13] W. Luo, P. Xing, Y. Cao, H. Yao, *et al.*, "Ura-net: Uncertainty-integrated anomaly perception and restoration attention network for unsupervised anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[14] P. Velickovic, W. Fedus, W. Hamilton, P. Liò, Y. Bengio, *et al.*, "Deep graph infomax," *stat*, 2018.

[15] C. Gallicchio and S. Scardapane, "Deep randomized neural networks," in *International neural network society big data and deep learning conference*, 2020.

[16] R. Panigrahi and S. Borah, "A detailed analysis of CICIDS2017 dataset for designing intrusion detection systems," *International Journal of Engineering and Advanced Technology*, 2018.

[17] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015.

[18] K. Fawagreh, M. Gaber, and E. Elyan, "Random forests: From early developments to recent advancements," *Systems Science & Control Engineering*, 2014.

[19] H. Altunay and Z. Albayrak, "A hybrid CNN+ LSTM-based intrusion detection system for industrial IoT networks," *Engineering Science and Technology, an International Journal*, 2023.

[20] L. Ruff, R. Vandermeulen, N. Goernitz, *et al.*, "Deep one-class classification," in *International conference on machine learning*, 2018.

[21] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on machine learning - ICML '05*, 2005.

[22] D. Hutchins, I. Schlag, Y. Wu, E. Dyer, *et al.*, "Block-recurrent transformers," in *Advances in neural information processing systems 35*, 2022.

[23] C. Barker, D. Bethell, and S. Gerasimou, "Robust adversarial quantification via conflict-aware evidential deep learning," in *International conference on learning representations*, 2026.

[24] P. Dong, Y. Xiao, C. Pun, F. Peng, *et al.*, "Harnessing transferable adversarial examples via multi-layer attention-guided spatial transformations," *IEEE Transactions on Reliability*, 2026.

[25] R. Qi, G. Liu, J. Zhang, and J. Hsiao, "Rethinking explainable AI: The gap between saliency-based explanation and user understanding for object detection models," *International Journal of Human-Computer Studies*, 2026.

[26] P. Biecek, "DALEX: Explainers for complex predictive models in r," *Journal of Machine Learning Research*, 2018.

Dr. Sonal Sharma, Currently serving as an **Associate Professor** in the Department of MCA at Lakshmi Narain College of Technology (LNCT), Bhopal. Holds an M. Tech in Computer Technology and Applications and an MCA**.** She has over 17 years of teaching experience in Computer Science. She received Ph.D. degree in Computer Science & Engineering from LNCT University, in 2025.

Dr. Virendra Kumar Tiwari is a Professor in the MCA Department at LNCT, Bhopal, with over 20 years of academic experience. His expertise includes Artificial Intelligence, Machine Learning, IoT, Cloud Computing, Cybersecurity, and Blockchain. He has published numerous research papers, authored books, secured patents, and actively contributes to research, innovation, and academic excellence.

Dr. Neelu Singh is an Associate Professor at Lakshmi Narain College of Technology (LNCT), Bhopal, with over 20 years of teaching experience. She holds a Ph.D. in 2023 in Computer Science and specializes in Data Mining, Machine Learning, Network Security, DBMS, and Programming. She has published research papers, authored books, and actively contributes to academic research and faculty development.

Prof. Seema Joshi is an MCA faculty member at Lakshmi Narain College of Technology (LNCT), Bhopal, with 2.5 years of teaching and 3.5 years of industry experience. Her expertise includes Programming Languages, DBMS, Web Technologies, Artificial Intelligence, and Cyber Security. She has published research papers and holds patents in Machine Learning.