Explanation Report for Additional New Material

Compared to the conference paper, our submission to TOPS has been updated and expanded as we detail in what follows. We confirm that the paper meets the requirements for an extended conference paper with at least 25% new material.

- We **added four brand new sections**: Section 2, Section 3, Section 4, and Section 8.
  - Section 2 introduces related work, outlining the concept, advantages and disadvantages of complete certification and incomplete certification, existing work on deterministic incomplete certification and probabilistic incomplete certification, and the differences between our approach and existing methods.
  - Section 3 introduces the preliminaries, formally explaining several key definitions of robustness certification based on randomized smoothing, including base classifier, robustness guarantee, smoothed classifier, multi-order information, and evasion attacks. In addition, two figures are added to illustrate the workflow of the deep learning-based network intrusion detector and the lp-bounded certified radius of the multi-class classifier on the input.
  - Section 4 provides the problem statement, describing the threat model considered in this paper, three research problems it aims to address, the solution directions specific to each problem, and the key challenges that we address.
  - Section 8 discusses the application scenarios of the proposed dimensional robustness certification approach and the importance of diversity in the types of norms used to constrain the certified robust regions.
- We **added five whole new subsections:** Subsection 6.5 to Section 6 (Experimental Setup), and Subsections 7.5, 7.6, 7.7, and 7.8 to Section 7 (Evaluation Results).
  - Subsection 6.5 outlines the construction methods and parameters for two threat models used in the evaluation: evasion attacks represented by PGD and EAD and natural corruptions represented by Latency and Packet Loss.
  - Subsection 7.5 provides additional comparisons of the certified and empirical robustness of network traffic classifiers against SOTA methods for fine-grained similar intrusion detection using the newly added dataset.
  - Subsection 7.6 presents a detailed evaluation of the dimensional certified radius of the MARS-defended network traffic classifier on all feature dimensions and analyzes the top 5 robust and top 5 sensitive traffic features.
  - Subsection 7.7 evaluates the impact of different smoothing distribution types on robustness certification, covering Gaussian, Laplace, and Uniform distributions aligned with l_2, l_1, and l_∞norms, respectively.
  - Subsection 7.8 compares the time overhead of various certified defense methods, using the average certification time per sample on each class as a measure of certification efficiency.
- **We re-wrote and extended the following sections:**
  - Abstract: We described the novelty and applicability of MARS and added a conclusion on the effectiveness of dimensional robustness certification.
  - Section 1 Introduction:
    - We introduced two new subsections, namely Overview of our method and Contributions, and added a contribution stating that the dimensional

certified radius reflects the fine-grained robustness differences across feature dimensions, consistent with the empirical evaluation findings.
- ✧ We added a detailed description of evasion attacks based on adversarial malicious traffic in the network traffic domain.
- ✧ We extended the discussion of the differences between certified defense and empirical defenses and the need for research.
- ✧ We revised the language expression of the entire section and improved the rigor of expression.

➢ Section 5 Design of MARS:
- ✧ We constructed Algorithm 1 and provided detailed steps for calculating the dimensional radius weight.
- ✧ We constructed Algorithm 2 and supplemented the detailed steps for multi-order information-based robustness certification.
- ✧ We added Figure 5 to illustrate the effect of the optimized smoothing distribution on the noise samples used for smoothing the base classifier.
- ✧ We extended Section 5.3 and introduced three new subsections (5.3.1, 5.3.2, and 5.3.3) to explain the strategy for aligning the smoothing distribution sampling region with the l_p certified region, adding Figure 6 to visualize the spatial distribution of noise sampling from different distributions.

➢ Section 6 Experimental Setup:
- ✧ We introduced a new Section 6 to detail the experimental setup, providing configuration specifics to enhance result reproducibility.
- ✧ We expanded the introduction of three existing certified defense methods (VRS, FRS, BARS) and explained the rationale for selecting them as baselines. Table 1 was added to highlight differences between the baseline defenses and our method in handling heterogeneous inputs, robustness guarantee diversity, and threat models.
- ✧ We elaborated on the principles of the two NIDS architectures (CADE and ACID) used for evaluation, demonstrating their applicability to diverse network traffic classification tasks.
- ✧ We expanded the description of the preprocessing steps for constructing sub-datasets and introduced a Similar-Intrusions dataset containing similar intrusion types for fine-grained intrusion detection. Details of the one-hot encoding process for categorical features were also provided.
- ✧ We added calculation details for mean certified radius, certified accuracy, robust accuracy, and clean accuracy, along with four additional metrics: recall, precision, false positive rate, and false negative rate.

➢ Section 7 Evaluation Results and Analysis:
- ✧ We redraw all bar figures in our experimental evaluation to add texture distinction to the legend.
- ✧ We rewrote the experimental analysis in Subsections 7.1.2, 7.2.2, 7.3.2, and 7.4.2 to provide a more comprehensive analysis of certified robustness and empirical robustness.

➢ Section 9 Conclusion: We added a summary of the core idea of the proposed method to search for a larger range of certified robust regions, and added text to

point out that the dimensional certified radius helps to locate the vulnerabilities of NIDS in feature learning and enhance the traffic feature-specific robustness.

- We **reorganized the paper structure**, and the number of method illustration figures and experimental charts is also more abundant.
  - Table 1: It outlines the updates of various attributes of our submission and the conference paper.
  - Table 2 (on the next page): It shows the updates in the paper structure of our submission and the conference paper. Newly added sections are highlighted in red, and rewritten and extended sections are highlighted in blue.

Table 1 Comparison of Paper Attributes

Item	Conference Paper	Paper Submitted to TOPS
Total Pages	8	32
Total Sections	5	9
Total References	42	56
Total Figures	7	16
Total Tables	1	6