IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 April 2024, accepted 17 June 2024, date of publication 1 July 2024, date of current version 23 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3421330

RESEARCH ARTICLE

Explaining Probabilistic Bayesian Neural Networks for Cybersecurity Intrusion Detection

TENGFEI YANG[1,2], YUANSONG QIAO[1], (Member, IEEE), AND BRIAN LEE[1], (Member, IEEE)

[1]Software Research Institute, Technological University of the Shannon: Midlands Midwest, Athlone, N37 HD68 Ireland
[2]Software College, Zhongyuan University of Technology, Zhengzhou 450007, China

Corresponding author: Tengfei Yang (yang_24@outlook.com)

This work was supported in part by the Technological University of the Shannon under the President Doctoral Scholarship, in 2021; and in part by the Horizon Europe Framework Program (HORIZON) through Resilmesh under Grant 101119681.

ABSTRACT The probabilistic Bayesian neural network(BNN) is good at providing trustworthy outcomes that is important, e.g. in intrusion detection. Due to the complex of probabilistic BNN, it is looks like a ‘‘black box’’. The explanation of its prediction is needed for improving its transparency. However, there is no explanatory method to explain the prediction of probabilistic BNN for the reason of uncertainty. For enhance the explainability of BNN model concerning uncertainty quantification, this paper proposes a Bayesian explanatory model that accounts for uncertainties inherent in Bayesian Autoencoder, encompassing both aleatory and epistemic uncertainties. Through global and local explanations, this Bayesian explanatory model is applied to intrusion detection scenarios. Fidelity and sensitivity analyses showcase that the proposed Bayesian explanatory model, which incorporates external uncertainty, effectively identifies key features and provides robust explanations.

INDEX TERMS Bayesian explanation, Bayesian autoencoder, uncertainty quantification, explainability, aleatoric and epistemic uncertainties.

I. INTRODUCTION

In the rapidly evolving landscape of cybersecurity, the deployment of probabilistic deep learning models has become increasingly prevalent for their unparalleled ability to discern complex patterns within vast datasets. Despite their efficacy, the opaqueness of these models presents a formidable challenge to understanding their decision-making processes. This can give rise to problems in critical applications where mistakes can be costly, including cybersecurity where it can contribute to the generation of excessive amounts of false alerts [1] by intrusion detection based IDS.

The reliability of intrusion detection results plays a critical role in determining the usability of the detection model. To address the deep learning model’s tendency towards overconfidence, enhancing trustworthiness involves two crucial aspects: quantifying uncertainty and providing explanations for model outputs [2]. The probabilistic Bayesian Neural Network (BNN) model, e.g. Bayesian Autoencoder (BAE) with uncertainty quantification, utilizes Bayes’ rule to yield probability outcomes, which is a widely adopted method [3] to get a trustworthy model by uncertainty quantification.

However, their inherent complexity often renders them as ‘‘black boxes’’, leaving cybersecurity practitioners and stakeholders in the dark as to their operation. This lack of explainability not only hinders the broader adoption of deep learning in cybersecurity but may also pose serious concerns regarding the potential introduction of biases or vulnerabilities that may go unnoticed. At the same time, curious analysts or experts don’t like to rely on model outputs without understanding the additional reasoning behind certain predictions that would earn the user’s trust and confidence.

For the black box issue, the need for eXplainable AI (XAI) methods for understanding and improving trust in AI models has arisen [4]. *Explainability* attempts to provide a human

The associate editor coordinating the review of this manuscript and approving it for publication was Luca Cassano.

interpretable reasoning for the model outcomes [5]. One such approach examines how each data feature influences the model's outcomes. This proves valuable in assisting experts in verifying the model's correctness and guiding them in making necessary improvements.

The estimation of uncertainty in the BNN model enhances confidence in the results. However, it is imperative to acknowledge and consider the impact of this uncertainty on the model's interpretation. This holds particular significance in critical domains, such as intrusion detection.

Several studies investigating explanations of BNN's predominantly employ deterministic explanations [6], [7], [8], [9] [10], and some use ensemble methods to quantify uncertainty in explanatory models [6], [7], [11].

However there has been been very few studies combining explainability with uncertainty analysis [2], [6], [7]. In addition, recent research on feature attribution has addressed the unreliability of salience maps when the test point is out-of-distribution [12].

As a result of this shortcoming we examine feature-based explainability for BNNs with uncertainty quantification (specifically for a BAE instance) whilst considering the impact of the uncertainty quantification on the model explanation. We therefore propose to derive a Bayesian method (*explanatory model* or interpretable model) [5] to attribute feature based explanations for the BAE with uncertainty quantification (*reference model* or original model). This method not only provides insights into the BAE with uncertainty quantification (BAE-UQ in short), but also give a Bayesian explanatory score with uncertainty to support further decision-making.

It is necessary at this stage to distinguish between two distinct issues, both falling under the umbrella of "explanatory uncertainty":

1) Explanatory External Uncertainty: the uncertainty in the reference model when this model is based on Bayesian methods, as e.g. in BNN. The uncertainty in the reference model not only influences the actual predictions (output by the reference model) but also introduces uncertainty into the explanation of the prediction results output by the explanatory model.
2) Explanatory Internal Uncertainty: the uncertainty in the explanatory model itself i.e. when the explanatory model is based on Bayesian methods. The consequent inherent uncertainty in the explanatory model introduces uncertainty into the explanatory scores.

In this paper, our focus is solely on the first category, referred to as explanatory external uncertainty. The structure of this research is shown in Fig. 1. The BNN model furnishes predictions in the form of anomaly scores, subsequently quantifying the uncertainties associated with these scores, encompassing both aleatoric and epistemic uncertainties. Through a holistic assessment of the anomaly score and its accompanying uncertainties, a Bayesian approach is employed (in the explanatory model) to furnish comprehensive explanations. Additionally, the external uncertainty of these explanations is estimated, further enhancing the explainability and reliability of the reference model's insights.

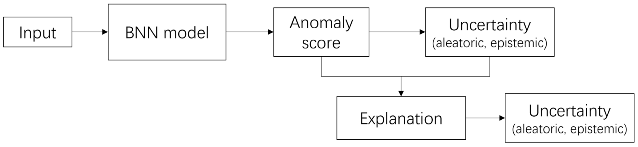
Input
BNN model
Anomaly score
Uncertainty (aleatoric, epistemic)
Explanation
Uncertainty (aleatoric, epistemic)

FIGURE 1. **The structure of this research.**

Moreover, the explanatory model can provide both *local* and *global* feature explainability. At a global level the goal is to identify the most relevant features for a given model across all the data instances whilst local explainability aims to identify the most relevant features for each each data instance separately [5]

Our contribution therefore is:

- An explanatory model is proposed for the BAE reference model concerning both aleatoric and epistemic uncertainties (BAE-UQ).
- The explanatory model uses Bayesian methods to produce a Bayesian explanatory score for both local and global model explanations.
- The Bayesian explanatory score considers the effect from the aleatoric and epistemic uncertainties of the BAE model, further provided interpretation with the external uncertainty quantification.
- This interpretable method combined with BAE-UQ is applied on the real intrusion detection dataset.

The remainder of the paper is structured as follows: Section II provides a literature review, while Section III delineates the interpretation method employed for BAE-UQ. Section IV presents experiments, covering both global and local explanations. Subsequently, Section V presents a thorough discussion of the results obtained. Finally, Section VI outlines the future directions and draws conclusions based on the findings presented.

II. LITERATURE REVIEW

Numerous methods have been proposed for explaining deep learning models, and one notable approach is Local Interpretable Model-Agnostic Explanations (LIME). Developed by Ribeiro et al. [13], LIME provides a model-agnostic methodology for generating locally faithful explanations for individual predictions. By systematically perturbing input instances and observing their impact on model outputs, LIME constructs interpretable surrogate models that approximate the behavior of the underlying deep learning model within a local context. This fine-grained explainability not only enhances the understanding of model decisions but also facilitates the identification of vulnerabilities. This, in turn, contributes to the robustness and trustworthiness of deep learning applications in the field of cybersecurity.

Bykov et al. [6] delved into the explanation of uncertainty in BNN, where uncertainty is treated as an anomaly score. In their work, Layer-wise Relevance Propagation (LRP) and integrated gradients serve as explanatory score methods within the ensemble BNN architecture. Importantly, the choice of the explanatory score method is flexible. The union and intersection explanation, in practical terms, translates BNN's predictive uncertainty into uncertainty associated with input features. This enriches XAI explanations, especially in the context of image data. In the study conducted by Clare et al. [7], uncertainty quantification in BNN is investigated using entropy. In this approach, the model prediction itself serves as a measure of uncertainty, and explanations are provided from both local and global perspectives to elucidate the results. The ensemble LRP and the ensemble SHAP (SHapley Additive exPlanations) are employed as two explanatory models. These models not only offer their respective on the explanatory value ranges for ocean and climate awareness in images but also provide a nuanced understanding of uncertainty levels through ensemble techniques.

The exploration of uncertainty in BNN models encompasses various methodologies. Depeweg et al. [8] investigated the sensitivity of aleatoric and epistemic uncertainties with respect to input features in BNNs with latent variables, utilizing a gradient-based approach.

In image analysis, Zintgraf et al. [9] generated smooth salience maps to visualize uncertainty in probabilistic models. Moving forward, Piironen et al. [14] employed a Bayesian model as a reference, utilizing Lasso and the entire elastic net family for feature selection to enhance accuracy. Peltola et al. [10] extended this method, combining KL-divergence with LIME (LIME-KL) for local interpretation of BNN predictive models. KL-divergence measures the difference between the predictive model's output and the explanatory model's output, acknowledging that in practice, the prediction is the mean of a distribution.

In an improved version, Afrabandpey et al. [11] employed a classification and regression tree as an explanatory model to offer both local and global explanations for Bayesian predictive models. Both studies highlight that aleatoric and epistemic uncertainties originating from the reference model can be captured in the interpretive model. The explanatory model is fitted to match the reference posterior predictive distribution, achieving aleatoric uncertainty by aligning with the reference model's posterior predictions and epistemic uncertainty by fitting the interpretive model to multiple posterior predictions.

A noteworthy probabilistic model serving as an explainable model has been formulated within a unified framework [5]. This framework is designed to achieve both global and local explainability for complex machine learning models, introducing the concept of Bayesian Importance of Features (BIF).

In this unified framework, a classification model is initially trained to learn the weights. Subsequently, a Bayesian explanatory model is linked to this classification model, with fixed weights. The input to the classification model is the product of the input data and the sample of the explanatory score. The entire model is then trained to obtain the Bayesian explanatory score, which adheres to a Dirichlet distribution. This score reflects the relative importance of each feature to the model output. The explanatory model can take the form of variables for global explanations or a neural network model for local explanations. The output of the neural network model serves as the hyperparameters for the posterior distribution of the explanatory scores. The experimental validation of this approach was conducted using a subset of the KDD'99 dataset.

III. METHOD

A. APPROACH

This research focuses on explaining probabilistic BNN through a Bayesian model. Following the framework established by [5], the BAE with uncertainty quantification (BAE-UQ) serves as the reference model, emphasizing high accuracy without the need to concern explainability. The quantification of uncertainty in the BAE encompasses both aleatoric and epistemic uncertainties [15]. The Bayesian model, following [11], is utilized as an explanatory model, employing KL-divergence to align the Bayesian model's predictions closely with those of the BAE-UQ. The weights in the Bayesian model follow a distribution, specifically the Dirichlet distribution as used in [11]. The posteriors of these weights constitute the Bayesian explanatory score, showcasing the relative contribution of each feature to the prediction.

We introduce a novel Bayesian explanatory score incorporating external uncertainty for probabilistic BNN. Aleatoric uncertainty, arising from unknown data acquisition factors, manifests itself in predictive distributions with noise. The output distribution of the explanatory model is designed to infinitely approximate the output distribution of the reference model. Through backward transfer, this output distribution of the explanatory model results in a distribution that aligns with the values of the explanatory scores. Consequently, the effect of aleatoric uncertainty on interpretation is reflected in the variation of explanatory score values. For epistemic uncertainty, commonly adopted in probabilistic BNNs, the ensemble method is utilized. Employing this approach, the external epistemic uncertainty of the Bayesian explanatory score is obtained. The Bayesian explanatory scores thus provide a foundation for expert judgment, considering external uncertainties for a comprehensive understanding.

Let $X = \{x_n\}_{n=1}^{N}$, represent a data set of size N, where $x_n = (x_{n1}, \ldots, x_{nd})^T$ is a D-dimensional feature vector, and $y_n \in \mathbb{R}$ is the target (either discrete or continuous). There is a highly predictive BAE-UQ (reference) model $f(x)$ fitted to the training data without explainability constraints.

The uncertainties include aleatoric uncertainty pertains to the inherent randomness inherent in an output and epistemic uncertainty arises from the variability of the parameters. In practical implementations of BAE models for uncertainty quantification, distinct methodologies are employed for each uncertainty. Specifically, aleatoric uncertainty is often

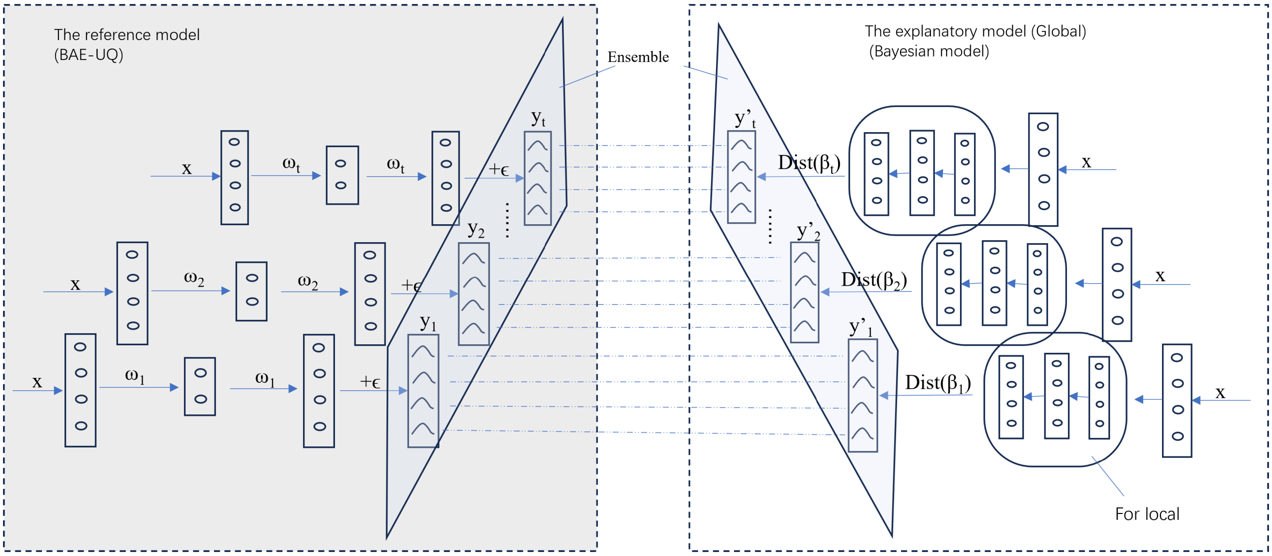

FIGURE 2. Bayesian explanatory model framework for ensemble BAE.

modeled by introducing noise ϵ to the data, the $\epsilon \sim \mathcal{N}(0, 1)$, the prediction $y_t \sim \mathcal{N}(f(x), \epsilon)$. Meanwhile, for capturing epistemic uncertainty, ensemble BAE models are commonly utilized, with parameters ω sampled from the trained posterior distribution of the encoder and the decoder. As depicted in Fig. 2, $(\omega 1, \omega 2, \ldots, \omega t) \sim Dist(\alpha)$, where α is the trained hyper-parameters. Functioning as a reference model, the BAE-UQ is regarded as a black box in the context of Bayesian explainability. In Bayesian explainability, each BAE-UQ prediction distribution is elucidated by a corresponding Bayesian explanatory model, wherein the parameters of the Bayesian explanatory model are represented as a distribution with hyperparameters β.

B. THE BAE AND UNCERTAINTY QUANTIFICATION

In the BAE model, Bayesian inference is applied to the parameters.The loss function [16], [17] shows as in equation (1).

$$-\int_{\Omega} q(\omega) \log p_x(y|x, \omega)\, d\omega + \mathrm{KL}[q(\omega) \,\|\, p(\omega)] \tag{1}$$

This first term represents the negative log-likelihood, where Ω is the space of possible parameters ω. The $q(\omega)$ is the variational distribution used for approximating the true posterior distribution. The $p(y|x, \omega)$ is the likelihood, representing the probability of the observed data y given the input x and parameters ω. The integral sums up the contribution of all possible parameter values. This second term represents the KL divergence between the posterior variational distribution $q(\omega)$ and the prior distribution $p(\omega)$. Due to the integral, the quantity is not easy to calculate. There are several approximating inference method, we choose Monte Carlo Dropout(MCD) [17] here. The model parameters are learned by maximizing the evidence lower bound(ELBO), as shown in the follow equation 2. A set of approximate posterior $\{\omega_m\}_{m=1}^{M}$ sampled from the posterior $q(\omega|x)$, and M is the number of samples.

$$\mathcal{L}_{\omega}(x) = \frac{1}{M} \sum_{m=1}^{M} \log p(y|x, \omega_m) - KL[q(\omega|x) \,\|\, p(\omega)] \tag{2}$$

The moment based predictive uncertainty quantification approach is used [15]. After the training phase, the hyperparameter α is optimized, and consequently, $\{\hat{\omega}_t\}_{t=1}^{T}$ is sampled from the distribution with the optimised hypterparamter α. Using Monte Carlo approximation, the variance $Var(y^*)$ of the predictive distribution for new data (x^*, y^*) during the testing phase is estimated as follows:

$$Var(y^*) \approx \frac{1}{T} \sum_{t=1}^{T} \left[\mathrm{diag}\{p(y^*|x^*, \hat{\omega}_t)\} - p(y^*|x^*, \hat{\omega}_t)^{\otimes 2} \right] \tag{3}$$

$$+ \frac{1}{T} \sum_{t=1}^{T} \left\{ p(y^*|x^*, \hat{\omega}_t) - \hat{p}(y^*|x^*) \right\}^{\otimes 2} \tag{4}$$

The $p(y^*|x^*, \hat{\omega}_t)$ is the predictive probability of y^* given x^* and the sample $\hat{\omega}_t$, $\mathrm{diag}\{p(y^*|x^*, \hat{\omega}_t)\}$ is a diagonal matrix with the elements of $p(y^*|x^*, \hat{\omega}_t)$ on the diagonal. $p(y^*|x^*, \hat{\omega}_t)^{\otimes 2}$ is the outer product of the predictive distribution with itself. Equation (3) captures the aleatoric uncertainty, and equation (4) captures the epistemic uncertainty. The $\hat{p}(y^*|x^*)$ is the expectation of the prediction.

C. THE BAYESIAN EXPLANATORY MODEL

Our objective is to identify a Bayesian explanatory model $g(x^*)$ that effectively captures the behavior of the reference model, characterized by the likelihood $p(y^* \mid x^*, \eta, g)$, where η represents the parameters of the explanatory model, also referred to as the Bayesian explanatory score, with η following a distribution represented as $Distribution(\beta)$.

We opt for the Dirichlet distribution as the distribution that governs the Bayesian explanatory score, thereby allowing the Bayesian explanatory score to illustrate the relative importance of each feature to the output [5].

The Bayesian explanatory score is a universal D-dimensional vector which is assigned to the set of input features in the dataset X. This vector quantifies the relevance or importance of each feature for the entire dataset, providing a global probabilistic explanation for each feature. Alternatively,the Bayesian explanatory score can be a matrix, each data instance x_n is allocated a distinct D-dimensional feature importance score. This approach offers a local probabilistic explanation for each record in the dataset.

The ideal explanatory model is the one that closely mirrors the predictions of the reference model while ensuring explainability. To measure the similarity in predictive behavior between the explanatory model and the reference model (probabilistic BNN), we calculate the KL divergence between their predictive distributions, as outlined in [11]. Additionally, we take into account the KL-divergence between the prior distribution and the posterior distribution of the Bayesian explanatory parameters is, as shown in equation (5).

$$\hat{\eta}_t = \arg\min_{\eta} \int \pi_{x^*}(z)\mathrm{KL}\left[p(y^* \mid z, \hat{\omega}_t, f) \parallel p(y^* \mid z, \eta, g)\right] dz \\ + \Phi(\eta) + \mathrm{KL}[q(\eta)||p(\eta)]. \qquad (5)$$

The expression involves the KL divergence, and $\pi_{x^*}(z)$ which is a probability distribution defining the local neighborhood around x^*, the data point for which the prediction is to be explained. Minimizing the KL divergence ensures that the interpretable model exhibits comparable predictive performance to the reference model. The Φ represents the penalty function for the complexity of the interpretable model, We also choose the non-zero Bayesian score as the basis of penalty.

To calculate the expectation in equation (5), a set of samples $\{z_s\}_{s=1}^{S}$ from $\pi_{x^*}(z)$ is drawn using Monte Carlo approximation. Minimising KL-divergence is equal to maximising the expected log-likelihood of the explanatory model $\log p({y^*}_s|z_s, \eta)$ over the posterior likelihood of the reference model $({y^*}_s|z_s, \hat{\omega})$ [14]. Combined to KL-divergence on the parameters, the Bayesian explanatory model obtained by maximizing ELBO:

$$\arg\max_{\eta} \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{{y^*}_s|z_s,\hat{\omega}_t}\left[\log p({y^*}_s|z_s, \eta)\right] - \Phi(\eta) \\ -KL[q(\eta)||p(\eta)] \qquad (6)$$

The final Bayesian explanatory score η' is the average of the Bayesian explanatory score for each sample prediction of the reference model.

$$\eta' = \frac{1}{T}\sum_{t=1}^{T} \hat{\eta}_t \qquad (7)$$

1) GLOBAL EXPLANATION

A probability vector serves as the vehicle for describing feature importance, providing a nuanced depiction of the relative weight of each feature. Importance is interpreted as the contribution of a feature to closely align with the prediction distribution of the reference model, denoted as η. We employ Dirichlet distribution as the prior distribution $\mathrm{Dir}(\beta_0)$. Both β and β_0 represent parameter vectors within the Dirichlet distribution. We maintain the value for the parameters β_0 as fixed constants, focusing our optimization efforts solely on the parameters β. The utilization of the Dirichlet distribution for both the posterior and the prior distributions facilitates the derivation of a closed-form KL-divergence in equation (6). Consequently, the objective function outlined becomes contingent on the Dirichlet parameters.

2) LOCAL EXPLANATION

Local explanations diverge from global ones in their assessment of feature importance, as they evaluate the importance of each feature for individual data instances. Unlike the global setting, which yields a single vector, the local context produces an importance matrix. Specifically, every data point x_n^* is assigned a vector that represents the feature importance for that specific data point. Consequently, a Bayesian importance matrix of size N × D is formed for the dataset X. In terms of parameterization, within the realm of local explains, each importance vector follows the posterior Dirichlet distribution with individual hyper-parameters β_n. A neural network is utilized to derive the hyper-parameters β_n. Following equation (8), the corresponding feature importance score η_n is drawn from the Dirichlet distribution with the parameter β_n for each data instance. Similar to the global case, we utilize the evidence lower bound; however, in the local scenario, the new objective function on η_n becomes reliant on the outputs of the neural network.

$$\arg\max_{\eta_n} \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{{y^*}_s|z_s,\hat{\omega}_t}\left[\log p({y^*}_s|z_s, \eta_n)\right] \\ - \Phi(\eta_n) - KL[q(\eta_n)||p(\eta_n)] \qquad (8)$$

In Bayesian inference within the Bayesian explanatory model, the model's output exhibits uncertainty, referred to as internal uncertainty. This internal uncertainty aligns with the common uncertainty quantification principles applied to BNN. It is important to note that the exploration of internal uncertainty falls within the broader scope of BAE uncertainty quantification and is not specifically addressed in this study. For those interested in delving into the details of internal uncertainty within BNN, pertinent literature on the subject is recommended for further reference [5].

3) EXTERNAL UNCERTAINTY

Regarding the uncertainty associated with the BAE-UQ, it's crucial to acknowledge that the model's output is not a single fixed value. The inherent uncertainties, namely aleatoric and epistemic uncertainties, contribute to the variability in the model's predictions. It's noteworthy that a single discrete explanatory model is insufficient for comprehensive explanations in the presence of such uncertainties.

In situations where both aleatoric and epistemic uncertainties are considered, traditional ensemble approaches may only effectively capture the epistemic uncertainty. This limitation emphasizes the need for more sophisticated methods that can handle both types of uncertainties for robust and accurate explanations.

The external epistemic uncertainty $ExUncer_{epis}$ of explanatory score is captured by ensemble the Bayesian explanatory score posterior distribution of each explanatory model $Dist(\beta_t)$. The variance of this ensemble group represents the external epistemic uncertainty, mirroring the approach used to capture epistemic uncertainty in the reference model. For practical implementation, the variance of the mean of each posterior distribution is computed. On the other hand, the external aleatoric uncertainty $ExUncer_{alea}$ of explanatory score originates from the variance of the prediction distribution of the reference model, as perceived by the explanatory model. The ideal prediction distribution of the explanatory model serves as an approximation of the reference model's predicted distribution. Through retrospective analysis, it becomes apparent that the variance of the prediction distribution of the explanatory model influences the variance of the parameters of the explanatory model $Var(Dist(\beta_t))$, thereby contributing to the overall external aleatoric uncertainty. As shown in follows:

$$ExUncer_{alea} = \frac{1}{T}\sum_{t=1}^{T} Var(Dist(\beta_t)) \tag{9}$$

$$\begin{aligned} ExUncer_{epis} &= Var(Dist(\beta_1), Dist(\beta_2), \ldots, Dist(\beta_t)) \quad (10)\\ &= Var(mean_{Dist(\beta_1)}, mean_{Dist(\beta_2)},\\ &\quad \ldots, mean_{Dist(\beta_t)}) \quad (11) \end{aligned}$$

Features exhibiting high epistemic uncertainty signal the need for vigilant monitoring or additional data collection to enhance model understanding and reduce uncertainty. On the other hand, features with elevated aleatoric uncertainty imply a connection with other unobserved or latent variables, underscoring the complexity associated with those particular features [8]. Identifying and addressing these uncertainties are crucial for improving model performance and reliability.

D. METRICS

To assess the effectiveness of the BAE model in intrusion detection, we rely on the AUC-ROC (Area Under the Receiver Operating Characteristic) metric, which evaluates the trade-off between the false positive rate and true positive rate. The AUC-ROC after rejected records with higher uncertainty than a threshold, denoted rejected-AUC, is used to quantify uncertainty in BAE-UQ.

Ensuring the transparency and explainability of the BAE model is imperative for their successful deployment in cybersecurity. Here, we use fidelity analysis, as demonstrated in [11], to measure the model explanation, by scrutinizing the alignment between a model's predictions and its interpretable counterpart. This analysis ensures that the model's behavior remains comprehensible and aligned with its intended objectives.

For the global explanation, sensitive analysis (leave-one-out) method is used for easy feature size selection [14]. This method facilitates the demonstration of the Bayesian explanatory score's efficacy by systematically evaluating the impact of individual or groups of features on model performance. Initially, features with the highest Bayesian explanatory scores are selected, and their performance is assessed using metrics such as AUC-ROC. This top-n features, which can take higher AUC-ROC than the whole features, demonstrates that the Bayesian explanatory score captured the correct importance. To enhance execution efficiency, feature selection is conducted incrementally in steps of 3, aiming for improved AUC-ROC values with fewer selected features.

The external uncertainty encompasses both aleatoric and epistemic uncertainties, which are aggregated to form the total uncertainty metric. To assess the impact of external uncertainty quantification on the Bayesian explanatory score, we progressively filter out features exhibiting higher uncertainty than a threshold and defer the evaluation of these features to domain experts. Subsequently, sensitivity analysis is conducted at each step to gauge the effect. To preserve the significance of the retained features, the rejection rate varies within the range of [1, D/2] with a step size of 7. By eliminating features with high uncertainty values, the remaining scores are deemed reliable, and the sensitivity analysis should reveal higher AUC-ROC values with fewer selected features compared to the unrejected scenario.

E. DECISION PROCESS

This section describes the final decision process according to uncertainties and the explanation, as in Fig. 3. From the start, after getting the prediction and the uncertainty in BAE-UQ, the Bayesian explanatory score of each prediction is calculated concerned aleatoric and epistemic uncertainties of the prediction. Simultaneously, the external uncertainty of the Bayesian explanatory score is calculated. At first, the uncertainty of the prediction needs to be estimated. If it is less than a threshold (threshold1) for BAE-UQ, the external uncertainty of the Bayesian explanatory score of each feature needs to be estimated in the next step. If the external uncertainty of a Bayesian score is larger than a threshold (threshold2), the score is rejected as it can not be trusted. The relative importance of the feature to the prediction is unclear. If the external uncertainty of Bayesian scores are less than the threshold2 for the explanation, a trustworthy prediction is provided, which means the previously calculated Bayesian explanatory score is validated. This also is the ideal outcome, named road①.

However, if the uncertainty of the prediction is larger than the threshold1, the prediction will normally be forwarded to

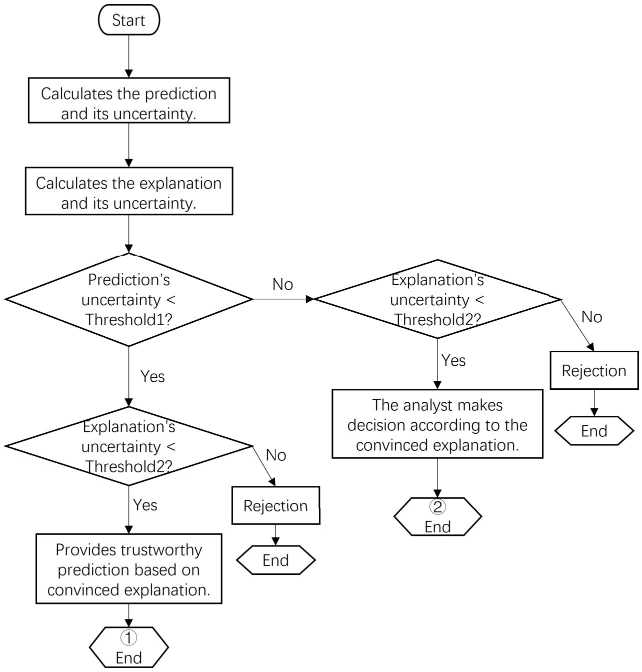
Start
Calculates the prediction and its uncertainty.
Calculates the explanation and its uncertainty.
Prediction's uncertainty < Threshold1?
No
Explanation's uncertainty < Threshold2?
No
Yes
Rejection
End
The analyst makes decision according to the convinced explanation.
②
End
Yes
Explanation's uncertainty < Threshold2?
No
Yes
Rejection
End
Provides trustworthy prediction based on convinced explanation.
①
End

FIGURE 3. The decision process.

an analyst to make a decision on its merit. In our process however the Bayesian explanatory score and the external uncertainty of each feature can assist the analyst to analyse. If the external uncertainty of a Bayesian explanatory score is larger than threshold2, the Bayesian explanation of this feature is suspect, the score is rejected as it can not be trusted.

If the external uncertainty of Bayesian scores are less than threshold2, the convinced Bayesian explanation is provided to the analyst, this is the road② in Fig. 3. In this way, the Bayesian explanatory score and the external uncertainty will aid the analyst in making quick decisions and reducing their workload.

IV. EXPERIMENT

This experiment leverages the UNSW-NB15 dataset and KDD'99 dataset to quantify both aleatoric and epistemic uncertainties in BAE, subsequently assessing global and local explanations in light of external uncertainty explanation. The specific interpretation process is illustrated in two case studies in UNSW-NB15.

UNSW-NB15 dataset [18] serves as a comprehensive repository of network activities, encompassing real benign network behaviors and synthetic attack scenarios. It categorised into nine sub-types, each representing a distinct cyber threat (refer to Table 1). The evaluation is conducted from the perspective of each attack type, where the benign class is partitioned into two subsets: 20% for testing and 80% for training. The 20% subset of normal data is then amalgamated with each attack type to form the respective test sets, enabling the assessment of the model's capability to detect each attack type.

TABLE 1. Contents on UNSW-NB15 dataset.

No.	Traffic Type	Number
1	Fuzzers	24246
2	Analysis	2677
3	Backdoors	2329
4	DoS	16353
5	Exploits	44525
6	Generic	215481
7	Reconnaissance	13987
8	Shellcode	1511
9	Worms	174
	Benign	2218761

KDD'99 [19], is a comprehensive and widely-used benchmark in the field of intrusion detection systems. This dataset was created by processing raw TCP dump data into a well-structured format, which includes a diverse range of network intrusions simulated in a military network environment. It consists of approximately 4.9 million records, each containing 41 features that capture various aspects of network traffic and host activities, as shown in Table 2. We divided the benign data into a training set (20%) and a test set (80%).

A. MODEL SETUP

A neural network is used to construct the encoder and the decoder in BAE model as the reference model with the size (128,64,32). The network architectures of BAE in MCD is shown in Table 3, which also set out the output shape, activation function and parameters of each layer. D is the original dimension, D equals 47 in UNSW-NB15 and equals 41 in KDD'99. In addition, the learning rate is configured as 0.0001 with mini-batch stochastic optimization method AMSGrad. The batch normalization method with momentum 0.95 and random Normal initializer is used to normalize the data before output. Each model was trained over 100 epochs. The number of samples, M=1 and the batch size equals 512 in the learning phase and the sample size T=5 in inference phase. An L2 regularizer with parameter 0.1 is used to regularize the weights and the bias in MCD, with initialization method He [20] under a Gaussian distribution, and with dropout rate 0.2. When modelling the aleatoric uncertainty in BAE, a diagonal multivariate Gaussian distribution is followed.

As for the Bayesian explanatory model, the batch sets to 512, epochs equals to 100, the sample size S is 5. In order to escape the local minima and saddle points and converge to the global optima, the annealing rate of KL-divergence on the parameters is set to 0.1 during training. For global interpretation, the learning rate is 0.6, the prior of the Bayesian explanatory model parameters follows a Dirichlet distribution with the concentration parameter 0.01. For local explanation, the learning rate is 0.0005, the prior distribution follows a Dirichlet distribution with parameter 0.5. The architecture of the neural network to derive the

TABLE 2. Contents on KDD'99 dataset.

No.	Traffic Type	Number
1	Probe	41102
2	DoS	3883370
3	U2R	52
4	R2L	1126
	Benign	972781

hyper-parameters in local explanation is shown in Table 4, which has 3 dense layers and 2 BatchNormalization layers.

In order to take aleatoric uncertainty and epistemic uncertainty into consideration and for ease of comparison, the sum of the two uncertainties is chosen as the basis of rejection both in the reference model and the explanatory model. Due to large numbers of combinations, results that maximise the accuracy are reported.

B. EVALUATION

The BAE-UQ is used as the reference model. Aleatoric uncertainty and epistemic uncertainty are quantified following the method in [15].

In this experiment, performance is evaluated from both global and local perspectives, focusing on the relative importance of each feature to the predicted outcomes of the reference model as indicated by the Bayesian explanatory score. For convenience, our Bayesian explanatory model, which considers KL-divergence and is referred to as BEM-KL, is compared to LIME and LIME-KL in terms of local explanation. The evaluation results are summarized in Table 5 and Table 6.

In the first part of the table 5, the global explanation, the AUC-ROC of the reference model on each attack type is firstly presented as baseline in the table, then the highest AUC-ROC of each attack type under top-n important features is shown. Remarkably, the AUC-ROC obtained with the top-n features surpasses the AUC-ROC of the reference model for each attack type. Additionally, the AUC-ROC under the external uncertainty quantification is shown in the fourth column. Compared to the AUC-ROC with top-n features, the higher AUC-ROC under the external uncertainty quantification on fewer number of features is obtained in each attack type. The fidelity and the standard deviation in global explanation are depicted in the next column, indicating that the global explanatory predictions closely align with the reference model predictions.

In the second part of the table 5, the local explanation is assessed by the fidelity and the standard deviation of LIME, LIME-KL and our proposed method. Specifically, we evaluate these metrics on records with minimized uncertainty for each attack type. A comparison across attack types reveals that LIME exhibits the highest fidelity, while our method demonstrates the lowest fidelity with a smaller variance.

The same performance in KDD'99 can be observed, as shown in Table 6. However, the fidelity of the global explanation is smaller than in UNSW-NB15, similar to the results within LIME-KL and BEM-KL. This indicates that the Bayesian explanatory model has better explainability in KDD'99 than in UNSW-NB15. Another noteworthy observation is that the U2R type in KDD'99 and the Generic type in UNSW-NB15 have worse AUC-ROC scores, less than 0.6, compared to other types. Fortunately, considering top-n features and uncertainty quantification can quickly enhance performance.

TABLE 3. Network architecture of BAE in MCD.

	Layer Type	Output Shape	Activation Function
Encoder	InputLayer	D	
	Dense	128	Relu
	Dropout	128	Relu
	Dense	64	Relu
	Dropout	64	Relu
	BatchNormal	64	
	Dense	32	Relu
Decoder	InputLayer	32	
	Dense	64	Relu
	Dropout	64	
	Dense	128	Relu
	Dropout	128	
	Dense	64	Sigmoid
	DistributionLambda	D	

TABLE 4. Network architecture of the local explanatory model.

Layer Type	Output Shape	Activation Function
InputLayer	D	
Dense	D	Relu
BatchNormal	D	
Dense	D	Relu
BatchNormal	D	
Dense	D	sigmoid

C. GLOBAL EXPLANATION CASE OF FUZZERS

This section shows the case of global explanation on the type of Fuzzers in UNSW-NB15. Fuzzers are a type of attack commonly employed to identify vulnerabilities in software by injecting invalid, unexpected, or random data. Given their potential to disrupt operations and compromise security, understanding and mitigating Fuzzers attack are paramount in safeguarding digital assets and maintaining the integrity of systems and networks.

Key features commonly analyzed to identify Fuzzers attack include unusual payloads, high input rate, protocol violation, systematic testing, abnormal traffic patterns, unexpected application behavior, input validation errors and exception handling triggers.

In the context of Fuzzers, external uncertainty quantification allows us to gauge the impact of external uncertainties to the model on its predictive performance. By examining the performance metrics under the external uncertainty, we can gain insights into how robust our model is in detecting Fuzzers attack amidst uncertainty.

TABLE 5. Measurements on UNSW-NB15 dataset.

	Global				Local Fidelity		
Attack Type	AUC (ref.)	Best AUC (top-n)	Best AUC (uncer.)	Fidelity	LIME	LIME-KL	BEM-KL
Fuzzers	0.912	0.935(28)	0.977(10)	0.2211(±6e-4)	0.3481(±8.1e-2)	0.2212(±4.1e-3)	0.2157(±9.6e-3)
Analysis	0.897	0.979(28)	0.986(4)	0.2212(±3.1e-7)	0.2338(±1.3e-2)	0.2253(±1.2e-4)	0.2134(±9.8e-3)
Backdoors	0.889	0.968(7)	0.984(4)	0.2209(±6.0e-7)	0.2418(±8.3e-2)	0.2335(±4.9e-3)	0.2263(±4.1e-5)
DoS	0.895	0.961(34)	0.963(4)	0.2208(±8.4e-4)	0.1558(±2.5e-2)	0.158(±5.6e-3)	0.1533(±6.1e-3)
Exploits	0.902	0.923(37)	0.961(19)	0.2214(±1.4e-3)	0.2657(±1.0e-1)	0.217(±3.9e-3)	0.2139(±1.0e-2)
Generic	0.425	0.619(13)	0.829(4)	0.2058(±2.4e-7)	0.4976(±1.7e-1)	0.2449(±1.2e-4)	0.2399(±1.1e-4)
Reconnaissance	0.901	0.935(22)	0.952(10)	0.2212(±9e-4)	0.2437(±3.1e-2)	0.2047(±1.1e-3)	0.2162(±6.8e-3)
Shellcode	0.896	0.945(25)	0.982(10)	0.2217(±4.1e-4)	0.3091(±1.7e-1)	0.1635(±4.4e-3)	0.1585(±1.2e-2)
Worms	0.952	0.98(19)	0.99(13)	0.2207(±4.4e-7)	0.2972(±1.5e-2)	0.2396(±6.3e-5)	0.2359(±2.8e-5)

TABLE 6. Measurements on KDD'99 dataset.

	Global				Local Fidelity		
Attack Type	AUC (ref.)	Best AUC (top-n)	Best AUC (uncer.)	Fidelity	LIME	LIME-KL	BEM-KL
Probe	0.931	0.932(40)	0.977(31)	0.0239(1.9e-6)	0.179(±7.6e-2)	0.0031(±6.1e-3)	0.0004(±2e-3)
DoS	0.971	0.992(22)	0.994(37)	0.0254(±2.5e-2)	0.5775(±4e-1)	0.0081(±6.8e-3)	0.0076(±2.7e-3)
U2R	0.537	0.539(37)	0.789(13)	0.0241(±2.6e-6)	0.592(±4.7e-1)	0.008(±1.1e-2)	0.0074(±3.2e-3)
R2L	0.615	0.882(34)	0.967(22)	0.024(±2.3e-6)	0.4597(±2.6e-1)	0.0179(±1e-4)	0.0128(±4.9e-3)

By mapping the distribution of uncertainty across predictions, a comprehensive view of model reliability is provided. Fig. 4 illustrates the reference model's uncertainty density map for the Fuzzers attack type. By analyzing this map, the epistemic uncertainty is less than 1e-5. We can discern regions where the model exhibits higher uncertainty, indicating areas where its predictions may be less reliable or more variable.

The global Bayesian explanatory score of each feature in Fuzzers attack is depicted in blue in the Fig. 5. Features are sorted by the Bayesian explanatory score that are represented in the form of the label of each bar. Each feature is positive important and the Bayesian explanatory scores are distributed in a step-wise fashion. Additionally, the external uncertainty associated with the Bayesian explanatory score is represented in orange and green in Fig. 5, falling in the range of [0,1]. For clarity, the external epistemic uncertainty values have been magnified by a factor of one hundred to enhance visibility and facilitate interpretation. As shown in this figure, epistemic uncertainty, less than 6e-4, is smaller than aleatoric uncertainty, even aleatoric uncertainty is less than 0.0047. This visualization enables a comprehensive understanding of both the relative importance of features captured by the Bayesian explanatory score and the associated uncertainty levels, providing valuable insights for decision-making and model evaluation in cybersecurity applications.

To assess the efficacy of external uncertainty quantification, a sensitivity analysis is conducted after iteratively rejecting features with high external uncertainty. The rejection is carried out in step sizes of 7, as illustrated in Fig. 6. The blue line represents the scenario with no rejected features, i.e., no uncertainty quantification, where the AUC-ROC is calculated using the top features determined by the Bayesian explanatory score. The highest AUC-ROC achieved is 0.935 with the top 28 features, as indicated in Table 5. As more features with high external uncertainty are rejected, the maximum AUC-ROC improves, requiring fewer important features. Notably, the maximum AUC-ROC of 0.977 is attained after rejecting 22 features with the highest uncertainty, using only the first 10 trusted features. This trend underscores the effectiveness of external uncertainty quantification in enhancing the model's performance and reducing the reliance on uncertain features.

Then, the global explanation of the Fuzzers attack with the maximum AUC-ROC is specified by means of features, taking into account both with and without external uncertainty quantification. As in Table 7, the first 28 features produces the highest AUC-ROC without concern uncertainty is listed, along with the interpretation that is provided by UNSW-NB15. There are 11 features that overlap with the feature subset found in [21]. For example, the most relative important feature 'proto' has the score about 0.054, which shows the transaction protocol. This feature is a key feature for detecting protocol violations and can signal a potential Fuzzers attack in the network traffic [22]. Feature 'dloss' and feature 'sloss' mean abnormal packets from destination and source, which signal abnormal traffic patterns or unexpected application behavior. Feature 'ct_ftp_cmd' shows a command in ftp session, which is the key of abnormal traffic patterns, input validation errors and exception handling triggers.

After removing 22 features with the highest external uncertainty, the first 10 features left together produces the best performance, including 'smeansz', 'dttl', 'ct_src_dport_ltm', 'ct_srv_src', 'dur', 'ct_dst_ltm', 'ct_src_ltm', 'sttl', 'Stime', and 'dbytes'. In summary the critical aspects for detecting Fuzzers attack in UNSW-NB15 are as follows:

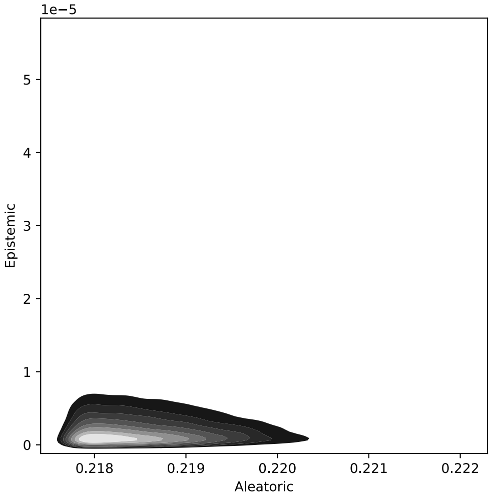

FIGURE 4. Uncertainty density of the Fuzzers.

- *Unusual packet lengths*: Fuzzing often involves sending packets with abnormal lengths, such as very short or very long packets ('smeanz' and 'dbytes'). Monitoring for packets with lengths significantly outside the normal range can indicate fuzzed traffic.
- *Unusual traffic patterns*: Analyzing the number of connections from the same source address ('ct_src_ltm'), or together with the destination port ('ct_src_dport_ltm'), or together with the same service ('ct_srv_src') at a time interval that may be indicative of high input rate or abnormal traffic patterns, as well as the number of connections to the same destination address at a time interval ('ct_dst_ltm'). Connections may also be of unusual laengths ('dur').
- *Protocol*: Protocol fuzzing typically targets specific network protocols or implementations. Monitoring the distribution of protocols ('proto') in the dataset and identifying unexpected or uncommon protocols being used could help detect protocol-specific fuzzing attack
- *Variations in TTL*:('sttl' as,'dttl'.) variations may be due to e.g. manipulating the TTL field directly to test for specific behaviors or vulnerabilities in network devices or protocols, packets that take different paths through the network help identify systemic testing from Fuzzers or unexpected application behavior.

D. LOCAL EXPLANATION CASE IN FUZZERS

This section delves into a case of local explanation within the Fuzzers category in UNSW-NB15, focusing on one specific record characterized by the minimal uncertainty (Fig. 7) in the reference model. The record with the minimal uncertainty indicates the most likely cause for the anomaly prediction to help validate the accuracy of the Bayesian explanatory score and the external uncertainty following route ① in Fig. 3.

As illustrated in Fig. 7, the Bayesian explanatory score for local explanations on records with minimal uncertainty follows a step-wise distribution. This suggests the absence of any particularly prominent features, aligning with global explanations. Features exhibiting external uncertainty above the threshold were removed, while the remaining features were analyzed for indications of a Fuzzer attack. The external aleatoric uncertainty and the external epistemic uncertainty are shown in orange and green respectively. The external epistemic uncertainty is also very small, about less than 1.4e-5, and enlarged 150 times for visual effect. Most features have very small external uncertainty less than 0.0028, except 'srcip', 'is_ftp_login', 'ct_src_ltm', 'Djit','ct_ftp_cmd', 'synack', and 'Sjit' that have relatively high external uncertainty.

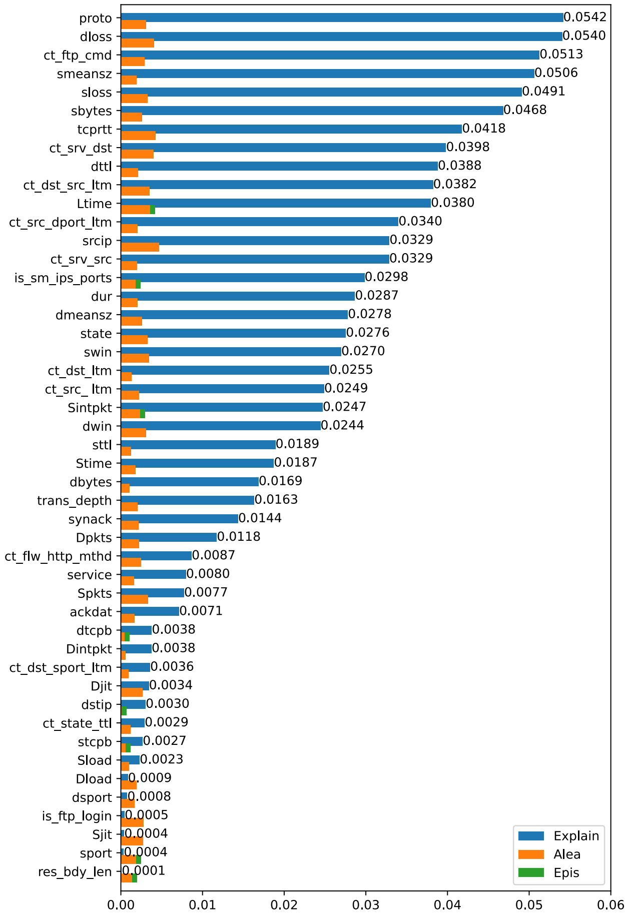

FIGURE 5. The Bayesian explanatory score with external uncertainty of global explanation for the type of Fuzzers.

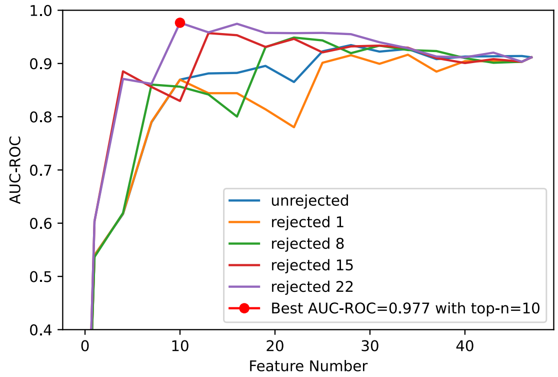

FIGURE 6. Sensitive analysis concerning the external uncertainty.

TABLE 7. Global Interpretation of key features.

No.	Feature name	Interpretation
1	proto	Transaction protocol.
2	dloss	Destination packets retransmitted or dropped.
3	ct_ftp_cmd	Count of flows that has a command in ftp session.
4	smeansz	Mean of the flow packet size transmitted by the src.
5	sloss	Source packets retransmitted or dropped.
6	sbytes	Source to destination transaction bytes.
7	tcprtt	TCP connection setup round-trip time, the sum of Synack and Ackdat.
8	ct_srv_dst	Count of connections that contain the same service and destination address in 100 connections according to the last time.
9	dttl	Destination to source time to live value.
10	ct_dst_src_ltm	Count of connections of the same source and the destination address in in 100 connections according to the last time.
11	Ltime	Record last time.
12	ct_src_dport_ltm	Count of connections of the same source address and the destination port in 100 connections according to the last time.
13	srcip	Source IP address.
14	ct_srv_src	Count of connections that contain the same service and source address in 100 connections according to the last time.
15	is_sm_ips_ports	If source and destination IP addresses equal and port numbers equal then, this variable takes value 1 else 0.
16	dur	Record total duration
17	dmeansz	Mean of the flow packet size transmitted by the dst.
18	state	Indicates to the state and its dependent protocol.
19	swin	Source TCP window advertisement value.
20	ct_dst_ltm	Count of connections of the same destination address in 100 connections according to the last time.
21	ct_src_ltm	Count of connections of the same source address in 100 connections according to the last time.
22	Sintpkt	Source interpacket arrival time (mSec).
23	dwin	Destination TCP window advertisement value
24	sttl	Source to destination time to live value.
25	Stime	Record start time.
26	dbytes	Destination to source transaction bytes.
27	trans_depth	Represents the pipelined depth into the connection of http request/response transaction.
28	synack	TCP connection setup time, the time between the SYN and the SYN_ACK packets.

This record is judged to be an abnormal after analysing features as in Table 8, which corresponds to the prediction (abnormal) and the real (abnormal). This table is summarized according to four risks on Fuzzers as follows.

1) Risk A: unusual packet lengths, i.e. high input rate
2) Risk B: protocol violation
3) Risk C: unusual payload
4) Risk D: abnormal traffic patterns

For each risk, specific features with value are listed in the second column, the description about the characteristics of the risk is the third column, the risk name is in the last column. The risks are sorted by the highest Bayesian score of features of each risk i.e. by descending likelihood of risk. Thus, the analyst can quickly make decision according to the top risks. For this record, the first (highest) risk is high input rate, which is an indicative of a Fuzzer attack. The second risk and the third risk et al. are supplemental and supportive of the prediction.

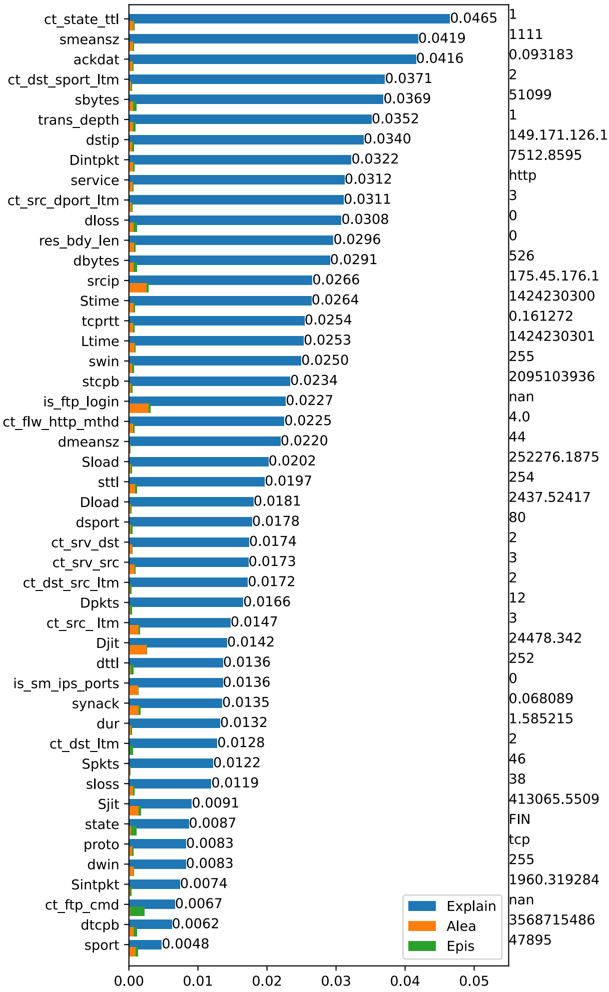

FIGURE 7. The Bayesian explanatory score with the external uncertainty of the record with the minimise reference model uncertainty.

V. DISCUSSION

We strive to enhance the reliability of intrusion detection systems by conducting thorough analysis and scrutiny within the UNSW-NB15 and the KDD'99 datasets. Our goal is to understand the inner workings of the models, improving their

TABLE 8. Local interpretation of key features (follow threshold).

No.	Feature (Value)	Interpretation	Risk

1	ct_state_ttl(1)
	smeansz(1111), dmeansz(44)
	sbytes(51099), dbytes (526)
	Dintpkt (7512.8595), Sintpkt (1960.319284)
		Count for each state according to specific range of values for source/destination time to live.
		Average packet size of Source (smeansz) and Destination (dmeansz).
		Transaction bytes of Source (Sbytes) and Destination (Dbytes).
		Interpacket arrival time(mSec) of destination(Dintpkt) and source(Sintpkt).
			A

2	ackdat(0.093183), synack(0.068089), tcprtt(0.161272)	TCP connection setup time. The time between SYN_ACK and the ACK (ackdat) is bigger than the time between the SYN and the SYN_ACK (synack).	B

3	ct_dst_sport_ltm (2), ct_src_dport_ltm (3)
	ct_srv_dst(2), ct_srv_src(3)
	ct_src_ltm(3), ct_dst_ltm(2)
	ct_dst_src_ltm (2)
		The count of connections that has same destination and source port (ct_dst_sport_ltm); the count of connections that has same source and destination port (ct_src_dport_ltm).
		The count of connections with the same source service(ct_srv_src); the count of connections with the same destination and service(ct_srv_dst).
		The count of connections with the same source(ct_src_ltm); the count of connections with the same destination(ct_dst_ltm).
		Count of connections of the same source and the destination address in a time interval.
			C;D

4	dloss(0), sloss(38)
	Dpkts(12), Spkts(46)
		Retransmitted or dropped packets of destination(dloss), and source(sloss)
		Packet count from destination to source (Dptks), and from source to destination (Spkts).
			C;D

5	Sload (252276.1875), Dload (2437.52417)	Bits per second of source (Sload) and destination (Dload).	C
	Predict:Abnormal	Real:Abnormal	

decision-making and reliability in real-world cybersecurity situations.

Our experiment highlights the effectiveness of global explanations in pinpointing important features for predictive model. The explanatory model shows a clear distinction in confident features when factoring in external uncertainty. For Fuzzers, the rejection threshold for external uncertainty assessment is vital; if too many features or crucial attributes are rejected based on external uncertainty, it affects performance. Therefore, carefully selecting the rejection threshold is crucial for optimizing sensitive analysis outcomes.

The case study on local explanations for records with minimized reference model uncertainty validates the effectiveness of external uncertainty quantification in facilitating confident interpretations for intrusion detection scenarios. Typically, records with higher uncertainty may yield either correct or incorrect predictions, requiring the intervention of analysts or experts for decision-making. External uncertainty quantification provides valuable insights into such records, enabling experts to mitigate uncertainties and make informed decisions.

Notably, the model performs worse on the Generic attack type in UNSW-NB15 and the U2R attack type in KDD'99 compared to other types. This variance may be due to the fact that Generic attacks have the highest number of records among attack types in UNSW-NB15, while U2R attacks have the fewest records in KDD'99. Therefore, the imbalance in record distribution needs to be analyzed for this model.

Additionally, it is essential to consider both computational complexity and scalability when applying our Bayesian explanatory model to real application. The model's complexity is influenced by the need to perform Bayesian inference, which inherently involves probabilistic computations that can be computationally intensive. This is particularly true when dealing with large datasets or high-dimensional data, where the number of parameters and operations increases exponentially. In the context of our Bayesian explanatory model, the incorporation of KL-divergence adds an additional layer of computational demands, as it requires the calculation of divergence between probability distributions, further contributing to the overall complexity.

Scalability, on the other hand, refers to the model's ability to handle increasing amounts of data without a significant degradation in performance. Our experiments indicate that while the Bayesian explanatory model provides robust and interpretable results, its scalability can be a challenge. Specifically, the Generic type in UNSW-NB15 has the most records, leading to a noticeable decrease in performance along with increased time and memory consumption. To mitigate these issues, optimization techniques such as parallel processing, efficient sampling methods, can be employed.

VI. FUTURE AND CONCLUSION

The explanation of the BAE-UQ serves to clarify the internal mechanisms and features driving predictions, enabling users to gain a deeper understanding of its functioning. Similar to uncertainty quantification, explainability enhances the transparency of the behavior of the BAE-UQ, thereby increasing trust of users, empowers users to make decision.

This research introduces a Bayesian model for globally and locally explaining the BAE-UQ, the aleatoric and epistemic uncertainties are transfer effect on the explanations as the external uncertainty. By integrating predictions and uncertainties of the BAE into the explanatory process, the experiment then transfers the Bayesian explanatory score to interpretable intrusion detection scene, providing decision-makers with a reliable and trustworthy basis for their decisions.

This method offers a reference for the explanation of the probabilistic BNN model. When the Bayesian model does not take external uncertainty into account, it is model-agnostic, allowing it to explain any model and derive the Bayesian explanatory score.

However, this method, which aligns with the probabilistic BNN uncertainty acquisition approach, is computationally demanding. Streamlining this process is the next challenge to address.

REFERENCES

[1] W. Alhakami, A. ALharbi, S. Bourouis, R. Alroobaea, and N. Bouguila, "Network anomaly intrusion detection using a nonparametric Bayesian approach and feature selection," *IEEE Access*, vol. 7, pp. 52181–52190, 2019.

[2] J. Antorán, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a CLUE: A method for explaining uncertainty estimates," 2020, *arXiv:2006.06848*.

[3] H. Wang and D.-Y. Yeung, "A survey on Bayesian deep learning," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–37, Sep. 2021.

[4] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805.

[5] K. Adamczewski, F. Harder, and M. Park, "Bayesian importance of features (BIF)," 2020, *arXiv:2010.13872*.

[6] K. Bykov, M. M.-C. Höhne, A. Creosteanu, K.-R. Müller, F. Klauschen, S. Nakajima, and M. Kloft, "Explaining Bayesian neural networks," 2021, *arXiv:2108.10346*.

[7] M. C. A. Clare, M. Sonnewald, R. Lguensat, J. Deshayes, and V. Balaji, "Explainable artificial intelligence for Bayesian neural networks: Toward trustworthy predictions of ocean dynamics," *J. Adv. Model. Earth Syst.*, vol. 14, no. 11, Nov. 2022, Art. no. e2022MS00316.

[8] S. Depeweg, J. M. Hernández-Lobato, S. Udluft, and T. Runkler, "Sensitivity analysis for predictive uncertainty in Bayesian neural networks," 2017, *arXiv:1712.03605*.

[9] L. M Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," 2017, *arXiv:1702.04595*.

[10] T. Peltola, "Local interpretable model-agnostic explanations of Bayesian predictive models via Kullback–Leibler projections," 2018, *arXiv:1810.02678*.

[11] H. Afrabandpey, T. Peltola, J. Piironen, A. Vehtari, and S. Kaski, "A decision-theoretic approach for model interpretability in Bayesian framework," *Mach. Learn.*, vol. 109, nos. 9–10, pp. 1855–1876, Sep. 2020.

[12] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim, "Debugging tests for model explanations," 2020, *arXiv:2011.05429*.

[13] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, ""Why should you trust my explanation?"" understanding uncertainty in LIME explanations," 2019, *arXiv:1904.12991*.

[14] J. Piironen, M. Paasiniemi, and A. Vehtari, "Projective inference in high-dimensional problems: Prediction and feature selection," *Electron. J. Statist.*, vol. 14, no. 1, pp. 2155–2197, 2020, doi: 10.1214/20-EJS1711.

[15] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, "Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation," *Comput. Statist. Data Anal.*, vol. 142, Feb. 2020, Art. no. 106816.

[16] Y. Gal, "Uncertainty in deep learning," Ph.D. thesis, Dept. Comput. Sci., Univ. Cambridge, Cambridge, U.K., 2016.

[17] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-on Bayesian neural networks—A tutorial for deep learning users," *IEEE Comput. Intell. Mag.*, vol. 17, no. 2, pp. 29–48, May 2022.

[18] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, Nov. 2015, pp. 1–6.

[19] S. Hettich and S. D. Bay, "The UCI KDD archive," Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 1999. [Online]. Available: http://kdd.ics.uci.edu

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[21] M. S. Al-Daweri, K. A. Z. Ariffin, S. Abdullah, and M. F. E. M. Senan, "An analysis of the KDD99 and UNSW-NB15 datasets for the intrusion detection system," *Symmetry*, vol. 12, no. 10, p. 1666, Oct. 2020.

[22] J. Li, B. Zhao, and C. Zhang, "Fuzzing: A survey," *Cybersecurity*, vol. 1, no. 1, pp. 1–13, Dec. 2018.

TENGFEI YANG received the M.Sc. degree in computer software and theory from Henan University of Technology, Zhengzhou, China, in 2011. She is currently pursuing the Ph.D. degree with the Software Research Institute (SRI), Technological University of the Shannon: Midlands Midwest working in the field of cyber security. Her research interests include network security and Bayesian deep learning.

YUANSONG QIAO (Member, IEEE) received the Ph.D. degree in computer applied technology from the Institute of Software, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently a Senior Research Fellow with the Software Research Institute (SRI), Technological University of the Shannon: Midlands Midwest, Ireland. He is a Science Foundation Ireland (SFI) Funded Investigator in the SFI CONFIRM Smart also Manufacturing Centre. His research interests include future internet architecture, blockchain systems, robotic control and coordination, and edge computing/intelligence. He is a member of IEEE (Robotics and Automation Society and Blockchain Community) and ACM (SIGCOMM and SIGAI).

BRIAN LEE (Member, IEEE) received the Ph.D. degree in computer science from the Trinity College Dublin, in 2004. He is currently the Director of the Software Research Institute (SRI), Technological University of the Shannon: Midlands Midwest, Ireland. He is also a Science Foundation Ireland (SFI) Funded Investigator in the SFI CONFIRM Smart Manufacturing Centre. His research interests include computer security (access control, network security, and security analytics) and programmable networking, and edge computing. He is a member of IEEE (Communications, Computer and Robotics and Automation Societies) and ACM.