Efficient Intrusion Detection Using Evidence Theory

Islam Debicha

Royal Military Academy
and Université Libre de Bruxelles
Brussels, Belgium
Email: debichasislam@gmail.com

Thibault Debatty

Royal Military Academy
Brussels, Belgium
Email: thibault.debatty@rma.ac.be

Wim Mees

Royal Military Academy
Brussels, Belgium
Email: wim.mees@rma.ac.be

Jean-Michel Dricot

Université Libre de Bruxelles
Brussels, Belgium
Email: jdricot@ulb.ac.be

***Abstract*—Intrusion Detection Systems (IDS) are now an essential element when it comes to securing computers and networks. Despite the huge research efforts done in the field, handling sources' reliability remains an open issue. To address this problem, this paper proposes a novel contextual discounting method based on sources' reliability and their distinguishing ability between normal and abnormal behavior. Dempster-Shafer theory, a general framework for reasoning under uncertainty, is used to construct an evidential classifier. The NSL-KDD dataset, a significantly revised and improved version of the existing KDD-CUP'99 dataset, provides the basis for assessing the performance of our new detection approach. While giving comparable results on the KDDTest+ dataset, our approach outperformed some other state-of-the-art methods on the KDDTest-21 dataset which is more challenging.**

Keywords–Intrusion detection; machine learning; evidence theory; contextual discounting.

I. Introduction

As computer network usage grows rapidly along with the significant increase in the number of applications running on it, the importance of network security is increasing. As dedicated tools designed to identify anomalies and attacks on the network, Intrusion Detection Systems (IDS) are becoming more valuable. Detection techniques based on anomalies and misuse have long been the principal subject of research in the field of intrusion detection [1].

Misuse-based IDSs operate quite similarly to most antivirus systems. Maintaining a signature database that could identify specific types of attacks and checking all incoming traffic against these signatures. Overall, this approach performs well, although it does not work properly when dealing with new attacks, or those that were specifically crafted to mismatch existing signatures.

On the other hand, anomaly-based IDSs operate generally on a baseline of normal activities and network traffic. This allows them to assess the current state of network traffic against this baseline so that abnormal patterns can be identified. While such an approach could be quite effective in detecting new attacks or those that have been intentionally crafted to evade IDSs, it can also result in a higher number of false positives compared to misuse-based IDSs.

Dempster-Shafer Theory (DST), also known as evidence theory [2] is one of the most versatile mathematical frameworks, extending Bayesian theory by (i) providing each source with the ability to integrate information at various scales of detail, thus addressing uncertainty; and (ii) offering a robust decision-making tool to make a consensus-based decision. This theory was later widely applied in several domains [3][4][5]. Regardless of this popularity, mass function generation and source reliability estimation remain an ongoing challenge.

Probabilistic frameworks for mass generation take advantage of the extensive research literature of the traditional probabilistic classifiers. These approaches usually represent the information associated with each attribute through Probability Density Functions (PDF), typically Gaussian [6][7]. Such densities are then transformed into beliefs that can subsequently be merged to form a joint decision. One can attribute masses to the compound hypotheses by subtracting the mass values related to the simple hypotheses involved [6] or by mixing the distributions associated with these hypotheses [7]. It should be noted that for most applications, Gaussian densities have been widely assumed due to their simplicity. Nevertheless, in the case where this assumption fails, the decision-making performance may be influenced considerably. More sophisticated approaches can be used to surmount this limitation by transforming data attributes into an equivalent normal space [8].

This paper offers a more effective way to overcome this disadvantage by constructing PDFs that are better suited to the original data histograms instead of projecting them into a new Gaussian-like space. On a more explicit level, a kernel smoothing estimation [9] is used on the training data to derive an approximate PDF for each data attribute and each simple hypothesis. These PDFs may be of any shape. Notably, they might be non-Gaussian. During the classification phase, a given datum is associated with a set of masses that are generated in

an elaborated way from the aforementioned densities. Using the proposed contextual discounting method, mass functions are then weakened differently depending on the ability of each source to discriminate between classes. Mass functions of the different sources are then merged to have a consensual mass function using a suitable fusion rule. A final decision is then deduced using the so-called "pignistic transform" [10].

The rest of this paper is organized as follows: Section II recalls the theoretical tools used in the proposed approach. Section III describes the NSL-KDD dataset. A description of Boosted PR-DS architecture is introduced in Section IV. Section V discusses the experimental results by comparing them with those of some previous studies using the NSL-KDD dataset. Final remarks and further suggestions for improvement are given in Section VI.

II. Related background

We succinctly outline some fundamentals of Dempster-Shafer theory, Parzen-Rosenblatt density estimation and contextual discounting.

A. Dempster-Shafer theory

Suppose that $\Omega = \{\omega_1, ..., \omega_K\}$, and $\mathcal{P}(\Omega) = \{A_1, ..., A_Q\}$ is its power set, where $Q = 2^K$. A defined mass function M ranging from $\mathcal{P}(\Omega)$ to $[0, 1]$ is named a "basic belief assignment" (*bba*) if $M(\emptyset) = 0$ and $\sum_{A \in \mathcal{P}(\Omega)} M(A) = 1$. A *bba* M therefore defines a "plausibility" function Pl ranging from $\mathcal{P}(\Omega)$ to $[0, 1]$ by $Pl(A) = \sum_{A \cap B \neq \emptyset} M(B)$, and a "credibility" function Cr ranging from $\mathcal{P}(\Omega)$ to $[0, 1]$ by $Cr(A) = \sum_{B \subset A} M(B)$. In addition, the two functions mentioned above are bound by $Pl(A) + Cr(A^c) = 1$. Moreover, a probability function p could be regarded as a particular case wherein $Pl = Cr = p$.

In case where two *bba*s M_1 and M_2 denote two elements of evidence, we can combine them together using the "Dempster-Shafer fusion" (DS fusion), which results in $M = M_1 \oplus M_2$ that is defined by:

$$M(A) = (M_1 \oplus M_2)(A) \propto \sum_{B_1 \cap B_2 = A} M_1(B_1) M_2(B_2) \quad (1)$$

Lastly, through Smets' technique[10], an evidential *bba* M can be converted into a probabilistic one, whereby every belief mass $M(A)$ is evenly distributed over all elements of A, resulting in the so-called "pignistic probability", Bet , given by:

$$Bet(\omega_i) = \sum_{\omega_i \in A \subseteq \Omega} \frac{M(A)}{|A|} \quad (2)$$

where $|A|$ is the number of elements of Ω in A.

It is worth mentioning that there are various evidential fusion rules in the literature that deal differently with the issue of conflicting sources [11][12][13].

B. Parzen-Rosenblatt density estimation

As a statistical tool, the Parzen-Rosenblatt window technique [14][15], otherwise known as kernel density estimation, is a way to smooth data by making population inferences based on a finite sample. This technique can be perceived as a non-parametric method to construct the PDF f, of an unknown form, linked to a random variable X. Suppose $(x_1, x_2, ..., x_N)$ an example of the realizations of such a random variable. The challenge is to estimate the f values at multiple points of interest. The smoothing of the kernel can then be seen as a generalization of the histogram smoothing where a window, of a predetermined shape, centered at every point is utilized to approximate the value of density at the given point. This is done by using the following estimator:

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \quad (3)$$

where $K(\cdot)$ is the kernel - a zero-mean non-negative function that integrates to one - and $h > 0$ is a smoothing parameter known as "kernel width". Furthermore, it is possible to use a variety of kernel functions like Uniform (Box), Gaussian (Normal), Triangle, Epanechnikov [16], Quartic (Biweight), Tricube [17], Triweight, Logistic, Quadratic [18], and others.

C. Discounting methods

Such methods can be used to estimate the weakening coefficients assigned to a source in order to correct its decision. These adjustments differ depending on whether it is a classic or contextual weakening.

1) Classical discounting: The weakening of mass functions makes it possible to model sources' reliability by introducing a coefficient α^s where for each source s we have:

$$\begin{cases} m'^s(A) = \alpha^s . m^s(A) & \forall A \in 2^\Omega, A \neq \Omega \\ m'^s(\Omega) = (1 - \alpha^s) + \alpha^s . m(\Omega) \end{cases} \quad (4)$$

α^s is the weakening coefficient of the s^{th} source. Among the classical weakening methods, we find [19] and [5].

2) Contextual discounting: The idea behind the contextual weakening is that the reliability of a source can vary depending on the truth of the object to be recognized (the context). For example, a sensor responsible for recognizing flying targets may be more or less able to discern certain types of aircraft. The method we propose belongs to this category and is described below.

Weakening using F-score: In this method, we evaluate the ability of each attribute (source) to classify elements belonging to different hypotheses -simple or composite-. This is done by considering each attribute separately to classify a new element. Using a cross-validation process, a confusion matrix is obtained. From this matrix, the "F-score" performance is calculated for all the hypotheses. These measures will be used as weakening coefficients and the equation 5 is applied to weaken the mass function of each source s.

$$\begin{cases} m'^s(A) = \alpha_A^s m^s(A) & A \in \left\{2^\Omega / \Omega\right\} \\ m'^s(\Omega) = m^s(\Omega) + \sum\limits_{A \in \{2^\Omega / \Omega\}} (1 - \alpha_A^s) m^s(A) \end{cases} \quad (5)$$

α_A^s is the weakening coefficient of hypothesis A for the s^{th} source.

III. NSL-KDD DATASET DESCRIPTION

In addition to the fact that attack patterns are constantly evolving and changing, the challenge in building a robust Network Intrusion Detection System (NIDS) is that a real-time pattern of network data consisting of both intrusions and normal traffic is out of reach. This is why many recent works are still using the NSL-KDD dataset to evaluate the performance of their approaches [20][21].

One of the most frequently used datasets for intrusion detection tests is the NSL-KDD dataset which was released in 2009 [22]. In addition to addressing efficiently redundant records' issue in the KDDCUP'99 dataset, NSL-KDD is designed by reducing the number of records in the training and test sets in a sophisticated manner to prevent the classifier from biasing towards frequent records.

There are three datasets within NSL-KDD. One for training which is KDDTrain+ and two for testing with an increasing difficulty respectively KDDTest+ and KDDTest-21, all of which having normal records as well as four distinct types of attack records, as shown in Table I. KDDTest-21 which is a subset of the KDDTest+ is designed to be a more challenging dataset by removing the often correctly classified records. For more details about how KDDTest-21 was conceived, the reader may refer to [22].

TABLE I. DIFFERENT CLASSES OF THE NSL-KDD DATASET.

	Normal	Dos	Probe	R2L	U2R
KDDTrain+	67343	45927	11656	995	52
KDDTest+	9711	7458	2421	2754	200
KDDTest-21	2152	4342	2402	2754	200

Each record has 41 attributes and a class label as well. These attributes are divided into basic features, content features, and traffic features. Attacks in the dataset are grouped into four categories based on their characteristics: DoS (denial of service attacks), Probe (Probing attacks), R2L (root-to-local attacks) and U2R (user-to-root attacks). Some specific types of attacks are included in the test set but are not included in the training set. This makes it possible to provide a more realistic testing ground.

IV. BOOSTED PR-DS

This section describes the theoretical basis of the proposed intrusion detection scheme called Boosted Parzen-Rosenblatt Dempster-Shafer (Boosted PR-DS). To do this, suppose we have a sample of N pre-tagged multiattribute data $(Z_1, ..., Z_N)$ where each datum $Z_n = (X_n, Y_n)$ with $X_n \in \Omega = \{\omega_1, ..., \omega_K\}$ being the tag, and $Y_n = (Y_n^1, ..., Y_n^P) \in \mathbb{R}^P$ being the P-attribute observation. The challenge is then to determine the tag of any new observation $Y_{n'}$.

As shown in Figure 1, we begin by briefly outlining the training process carried out on the pre-tagged data sample $(Z_1, ..., Z_N)$. Next, we illustrate the way our approach assigns a new observation $Y_{n'}$ to one of the K classes (tags).

A. Training phase

Consider the pre-tagged multi-attribute data above $(Z_1, ..., Z_N)$. Under our Boosted PR-DS scheme, the training phase involves two steps. The first is model adjustment which consists of determining the optimal kernel and fusion rule for the data along with the computing of weakening coefficients for each hypothesis. The second step is density estimation where the previously chosen kernel is used to estimate the Probability Density Functions (PDF) of each class for all attributes.

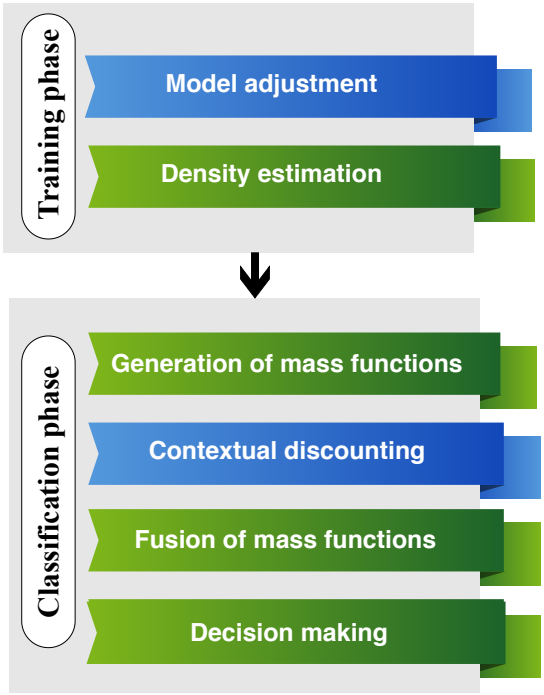


Figure 1. Proposed Boosted PR-DS framework

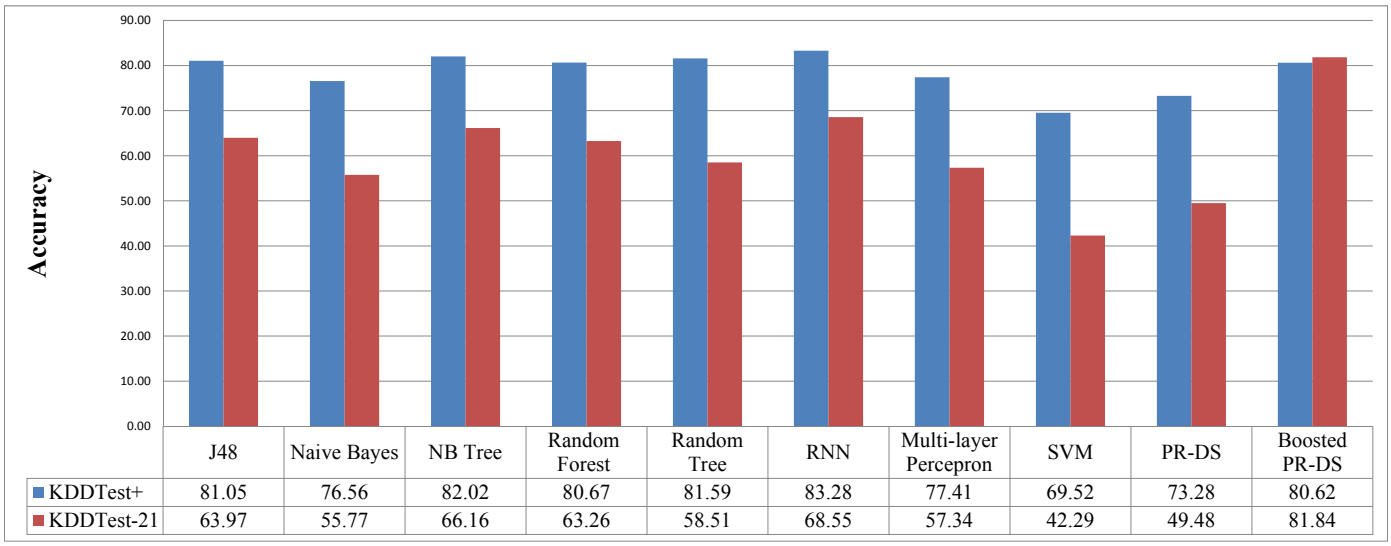
1) Model adjustment: In the first step, while changing kernels and fusion rules, basic PR-DS is used in a cross-validation process on the training data. The kernel and fusion rule giving the highest accuracy are then selected. To compute the weakening coefficients, we propose to use the F-score measures obtained from classifying each attribute (taken alone) as explained in paragraph Section II-C2.

2) Density estimation: In this step, we use the kernel chosen during the previous step to estimate densities using the Parzen-Rossenblatt method as described in Section II-B instead of considering that they follow a normal distribution as in the classical case. We thus obtain, for each class $\omega_k \in \Omega$ and for each attribute p $(1 < p < P)$, a Parzen-Rosenblatt density $\hat{f}_k^p$.

Eventually, in addition to the estimated densities, the trained model includes the weakening coefficients and the best-fit fusion rule.

B. Classification phase

Given a new observation $Y_{n'}$, a mass function M^p for each attribute is constructed based on the estimated densities. The proposed contextual discounting mechanism is then applied using the previously calculated weakening coefficients. Subsequently, the weakened mass functions are combined to obtain a consensual report M. The final decision is made using the so-called Pignistic Transform. In what follows, we describe these different steps.



	J48	Naive Bayes	NB Tree	Random Forest	Random Tree	RNN	Multi-layer Percepron	SVM	PR-DS	Boosted PR-DS
KDDTest+	81.05	76.56	82.02	80.67	81.59	83.28	77.41	69.52	73.28	80.62
KDDTest-21	63.97	55.77	66.16	63.26	58.51	68.55	57.34	42.29	49.48	81.84

Figure 2. Performance of Boosted PR-DS and the other models on KDDTest+ and KDDTest-21.

1) Generation of mass function: In order to determine the mass $\mathcal{M}^p$ assigned to the attribute p, we will consider the rank function δ_p which is defined from $\{1, .., K\}$ to Ω so that $\delta_p(k)$ is the $k-$ranked element of Ω in terms of $\hat{f}^p$, i.e. $\hat{f}^p_{\delta_p(1)}(Y^p_{n'}) \leq \hat{f}^p_{\delta_p(2)}(Y^p_{n'}) \leq ... \leq \hat{f}^p_{\delta_p(K)}(Y^p_{n'})$. Then, $\mathcal{M}^p$ is determined as follows:

$$\begin{cases} \mathcal{M}^p(\Omega) \propto \hat{f}^p_{\delta_p(1)}(Y^p_{n'}) \\ \mathcal{M}^p(\{\omega_{\delta_p(k)}, ..., \omega_{\delta_p(K)}\}) \propto \hat{f}^p_{\delta_p(k)}(Y^p_{n'}) - \hat{f}^p_{\delta_p(k-1)}(Y^p_{n'}) \end{cases} \tag{6}$$

2) Contextual discounting: To fine-tune the ultimate mass assigned to the p attribute, a weakening process based on the proposed contextual discounting mechanism mentioned in paragraph II-C2 is applied.

3) Fusion of mass functions: Mass Functions assigned to different attributes are then merged into a single consensus mass $M = \bigoplus_{p=1}^{P} M^p$ using the fusion rule selected on the training phase.

4) Decision making: The final decision is made based on the Pignistic transformation of M:

$$\hat{X}_{n'} = \arg\max_{\omega_k} \sum_{A \ni \omega_k} \frac{M(A)}{|A|} \tag{7}$$

It is worth noting that the novelty of Boosted PR-DS with respect to those using similar architectures is based on the steps of model adjustment, generation of mass function, and contextual discounting.

V. Experimental results

To assess the performance of the proposed boosted PR-DS method, experimental tests are conducted on the NSL-KDD dataset containing two test sets of increasing difficulty, KDDTest+ and KDDTest-21 respectively, as described in Section III.

A comparative analysis is made with regard to nine methods: J48 decision tree learning [23], Naive Bayes [24], NBTree[25], Random Forest [26], Random Tree [27], Multi-layer Perceptron [28], Support Vector Machine (SVM) [29], and Recurrent Neural Networks (RNN) [21], Parzen-Rosenblatt Dempster-Shafer (PR-DS) [30].

While giving a comparable accuracy on the KDDTest+ dataset, Boosted PR-DS outperforms the other state-of-the-art methods on the KDDTest-21 testing set as shown in Figure 2. This is mainly due to taking the estimated reliability into account by using the contextual discounting mechanism along with adjusting the model by selecting the most suitable kernel and fusion rule for a given training dataset.

To demonstrate the effect of kernel selection, we assess our approach on the KDDTest-21 dataset by changing the kernel each time, while maintaining the other parameters. Figure 3 shows that three kernels at least are getting better results than the Normal kernel which confirms the relevance of choosing an adapted kernel to suitably constructing our densities instead of using the normality assumption.

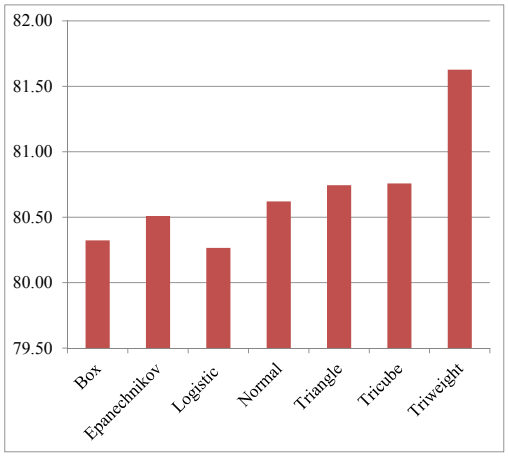


Figure 3. Boosted PR-DS on the KDDTest-21 dataset using different kernels

VI. Conclusion and future work

As a conclusion, we can consider Boosted PR-DS as a combination of multiple classifiers where each attribute (source) is a classifier. By using contextual discounting, one may prioritize the decision of an individual classifier regarding those classes in which its accuracy was high in the training phase and be doubtful regarding those classes it did not classify well. Furthermore, Boosted PR-DS choose a suitable fusion rule to take advantage of each individual classifier's knowledge to achieve a consensus decision. Experimental results validate the interest of this approach with respect to other state-of-the-art intrusion detection models. As a possible future direction, it would be interesting to consider handling conflicting sources with a more sophisticated fusion rule.

References

[1] J. Andress, The basics of information security: understanding the fundamentals of InfoSec in theory and practice. Syngress, 2014.

[2] G. Shafer, A Mathematical Theory of Evidence. Princeton: Princeton University Press, 1976.

[3] R. W. Jones, A. Lowe, and M. J. Harrison, "A framework for intelligent medical diagnosis using the theory of evidence," Knowledge-Based Systems, vol. 15, no. 1, 2002, pp. 77–84.

[4] M. E. Y. Boudaren, L. An, and W. Pieczynski, "Dempster–Shafer fusion of evidential pairwise Markov fields," International Journal of Approximate Reasoning, vol. 74, 2016, pp. 13–29.

[5] H. Guo, W. Shi, and Y. Deng, "Evaluating sensor reliability in classification problems based on evidence theory," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 5, 2006, pp. 970–981.

[6] F. Salzenstein and A.-O. Boudraa, "Unsupervised multisensor data fusion approach," in Signal Processing and its Applications, Sixth International, Symposium on. 2001, vol. 1. IEEE, 2001, pp. 152–155.

[7] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski, "Multisensor image segmentation using Dempster—Shafer fusion in Markov fields context," Geoscience and Remote Sensing, IEEE Transactions on, vol. 39, no. 8, 2001, pp. 1789–1798.

[8] P. Xu, Y. Deng, X. Su, and S. Mahadevan, "A new method to determine basic probability assignment from training data," Knowledge-Based Systems, vol. 46, 2013, pp. 69–80.

[9] M. P. Wand and M. C. Jones, Kernel smoothing. Crc Press, 1994.

[10] P. Smets and R. Kennes, "The transferable belief model," Artificial intelligence, vol. 66, no. 2, 1994, pp. 191–234.

[11] D. Dubois and H. Prade, "Representation and combination of uncertainty with belief functions and possibility measures," Computational intelligence, vol. 4, no. 3, 1988, pp. 244–264.

[12] P. Smets, "The combination of evidence in the transferable belief model," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 5, May 1990, pp. 447–458.

[13] F. Sebbak, F. Benhammadi, S. Bouznad, A. Chibani, and Y. Amirat, "An evidential fusion rule for ambient intelligence for activity recognition," in International Conference on Belief Functions. Springer, 2014, pp. 356–364.

[14] E. Parzen, "On estimation of a probability density function and mode," The annals of mathematical statistics, vol. 33, no. 3, 1962, pp. 1065–1076.

[15] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," The Annals of Mathematical Statistics, 1956, pp. 832–837.

[16] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," Theory of Probability & Its Applications, vol. 14, no. 1, 1969, pp. 153–158.

[17] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statistician, vol. 46, no. 3, 1992, pp. 175–185.

[18] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," Journal of the American statistical association, vol. 83, no. 403, 1988, pp. 596–610.

[19] Z. Elouedi, K. Mellouli, and P. Smets, "Assessing sensor reliability for multisensor data fusion within the transferable belief model," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, no. 1, 2004, pp. 782–787.

[20] S. Gurung, M. K. Ghose, and A. Subedi, "Deep learning approach on network intrusion detection system using nsl-kdd dataset," International Journal of Computer Network and Information Security (IJCNIS), vol. 11, no. 3, 2019, pp. 8–14.

[21] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," Ieee Access, vol. 5, 2017, pp. 21 954–21 961.

[22] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009, pp. 1–6.

[23] J. R. Quinlan, C4. 5: programs for machine learning. Elsevier, 2014.

[24] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[25] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." in Kdd, vol. 96. Citeseer, 1996, pp. 202–207.

[26] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, 2001, pp. 5–32.

[27] D. Aldous, "The continuum random tree. i," The Annals of Probability, 1991, pp. 1–28.

[28] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a bayes optimal discriminant function," IEEE Transactions on Neural Networks, vol. 1, no. 4, 1990, pp. 296–298.

[29] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, 2011, p. 27.

[30] A. Hamache, M. E. Y. Boudaren, H. Boukersoul, I. Debicha, H. Sadouk, R. Zibani, A. Habbouchi, and O. Merouani, "Uncertainty-aware parzen-rosenblatt classifier for multiattribute data," in International Conference on Belief Functions. Springer, 2018, pp. 103–111.