

# LEMMA: A Novel Feature Engineering Method for Intrusion Detection in IoT Systems

Ali Ghubaish, *Graduate Student Member, IEEE*, Zebo Yang, *Graduate Student Member, IEEE*,  
Aiman Erbad, *Senior Member, IEEE*, and Raj Jain, *Life Fellow, IEEE*

**Abstract**—Intrusion detection systems (IDS) for the Internet of Things (IoT) systems can use AI-based models to ensure secure communications. IoT systems tend to have many connected devices producing massive amounts of data with high dimensionality, which requires complex models. Complex models have notorious problems such as overfitting, low interpretability, and high computational complexity. Adding model complexity penalty (i.e., regularization) can ease overfitting, but it barely helps interpretability and computational efficiency. Feature engineering can solve these issues; hence, it has become critical for IDS in large-scale IoT systems to reduce the size and dimensionality of data, resulting in less complex models with excellent performance, smaller data storage, and fast detection. This paper proposes a new feature engineering method called LEMMA (Light feature Engineering based on the Mean Decrease in Accuracy). LEMMA applies exponential decay and an optional sensitivity factor to select and create the most informative features. The proposed method has been evaluated and compared to other feature engineering methods using three IoT datasets and four AI/ML models. The results show that LEMMA improves the  $F_1$  score performance of all the IDS models by an average of 34% and reduces the average training and detection times in most cases.

**Index Terms**—Feature engineering, Feature reduction, Feature selection, Internet of Things, IoT, Intrusion Detection Systems, IDS, Mean Decrease in Accuracy, MDA, Permutation feature importance

## I. INTRODUCTION

INTERNET of Things (IoT) plays a significant role in 21st-century technologies. According to Schiller et al. [1], the number of IoT devices is expected to jump from 31 billion in 2022 to 75 billion in 2025. Even with COVID-19 affecting supply and demand, resulting in a global chip shortage, these numbers were only slightly less, from 11.7 to 11.3, than forecast in 2020. This is because it integrates into different application domains of our daily lives, from Industrial IoT (IIoT) to critical Internet of Medical Things (IoMT). This increase in IoT systems will be accelerated in the 5G era as the network infrastructure enables large-scale cellular IoT (e.g., massive IoT). Due to the vast number of IoT devices in most IoT networks, these networks generate massive high-dimensional datasets, causing feature explosion [2].

Ali Ghubaish is with Washington University in Saint Louis, St. Louis, MO 63130 USA (email: aghubaish@wustl.edu).

Zebo Yang is with Washington University in Saint Louis, St. Louis, MO 63130 USA (email: zebo@wustl.edu).

Aiman Erbad is with College of Science and Engineering, Hamad Bin Khalifa University, Qatar (email: aerbad@hbku.edu.qa).

Raj Jain is with Washington University in Saint Louis, St. Louis, MO 63130 USA (email: jain@wustl.edu).

Manuscript received October 10, 2022.

Securing IoT systems is crucial due to the sensitive nature of the data produced by these sensors and devices (e.g., the medical data generated by IoMT systems). One of the solutions to secure these systems is by using large-scale intrusion detection systems (IDS), which can handle the massive and high-dimensional data generated by these systems [3], [4]. However, fitting these colossal datasets requires complex machine learning (ML) models, which brings challenges of overfitting, low interpretability, and high computational complexity. Besides, with high dimensional data, ML models for IDS require powerful devices and significant training time to detect attacks accurately. Thus, it is non-trivial to avoid the high complexity of data. Reducing complexity in data before modeling has been a hot topic in recent years [5]. Feature engineering is one popular method of such techniques. Feature engineering methods help select the best features for these models, expediting the processes of finding the optimal hyperparameters for the IDS models. We use “feature engineering” and “feature reduction” interchangeably in the rest of this paper to describe the methods that reduce the datasets’ dimensionality.

Most informative features can be selected using feature engineering or dimension reduction techniques, such as feature selection and feature extraction. Feature selection techniques help simplify complex models by reducing the dimensionality of the dataset (number of features), which avoids over-fitting and results in less training time and storage space. Only the most important features are retained after feature selection. On the other hand, feature extraction techniques, such as principal component analysis (PCA), create new features that preserve the data’s variance based on existing features.

Feature selection techniques are divided into four categories: filter, wrapper, embedded, and hybrid [6]. Filter methods are fast, but they may fail to select the most informative features and thus lead to low accuracy in ML models. Wrapper methods, like recursive feature elimination (RFE) and forward feature selection (FFS), are effective in selecting informative features but are slow and susceptible to overfitting. Embedded methods, such as the mean decrease in impurity (MDI) and the mean decrease in accuracy (MDA), provide a tradeoff between accuracy and speed, thus providing balanced results between filter and wrapper methods. Finally, hybrid methods are a mix of two or more of these methods (e.g., [7], [8]) and feature extraction such as PCA (e.g., [9]). However, these methods are usually designed for specific datasets or models. More about hybrid methods can be found in [10].

This paper proposes a new feature engineering method

for supervised ML-based IDS in IoT systems called Light feature Engineering based on the Mean Decrease in Accuracy (LEMMA). It takes the benefits of both feature selection and extraction methods and reduces their drawbacks. LEMMA is based on embedded methods and achieves a better tradeoff between performance and speed.

Our method consists of two parts: 1) creating a list of the most informative features using the MDA method 2) creating a new feature from the first feature (the most informative one) in that list. The new feature is created using the weighted exponential decay formula (WEDF) technique. In addition, in cases where the most informative feature is categorical, we utilize the Sensitivity Factor (SF) to complement the WEDF method for creating a new feature. This happens, for example, when most attacks are passive, e.g., sniffing. WEDF and SF optimize the relationship between the values in the most informative feature and the samples' classes, as shown in the Evaluation section of this paper.

LEMMA is a general feature engineering method using AI-based models for the supervised ML-based IDS in IoT systems. We demonstrate the effectiveness of our method by using three different datasets and comparing three different ML models using three different metrics. The evaluation results show the outstanding performance of LEMMA in IDS, with high accuracy and low detection time. The main contributions of this paper can be summarized as follows:

- We present LEMMA, a novel feature engineering method designed for supervised ML-based IDS in IoT.
- We show that the WEDF, when added to MDA, significantly improves the performance.
- We develop an add-on technique, SF, to enhance performance in cases where the most informative feature is categorical, which happens when most attacks are passive, e.g., sniffing.
- We have designed LEMMA to be compatible with a wide range of supervised ML models, such as random forest (RF), allowing straightforward implementation without hyperparameter tuning.
- We evaluate the performance of our method using three different datasets of different sizes, four different AI/ML models, and three different metrics – a total of 36 combinations.
- We empirically show that LEMMA improves the IDS performance by 34% in all cases and significantly reduces training and detection times in most cases.

The remainder of this article is organized as follows. A brief background of the commonly used feature engineering methods and the related work is provided in Sections II and III, respectively. In Section IV, we present our proposed method. The experimental methodology and results are shown in Sections V and VI. Finally, we conclude this paper in Section VII.

## II. BACKGROUND

This section presents a background of the most commonly used feature engineering methods.

### A. Categories of Feature Selection Techniques

The difference between the four common feature selection techniques – Filter, Wrapper, Embedded, and Hybrid – is briefly explained in this subsection. More detailed information can be found in [6] and [10].

**1. Filter:** In this technique, features are sorted based on their relevance. Then, a threshold is applied to select the features that have strong relevance. This results in a fast selection but may lead to low accuracy if the dataset distribution is not uniform and the features are highly correlated. The correlation coefficient method is an example of this technique. It measures the linear relationship between the features and selects the features with a correlation below a specific threshold.

**2. Wrapper:** In this method, the features are selected by measuring the performance improvement for an ML model using a subset of the features. The subset with the highest improvement in the ML model is selected. This technique effectively selects informative features but is very slow since it is computationally expensive, and the complexity increases as the number of features increases. For example, the RFE method uses all the features at the beginning, recursively removes them, and then sorts them by their incremental improvement.

**3. Embedded:** Embedded techniques combine the advantages of filter and wrapper techniques by embedding feature selection within the ML model itself. However, this makes it less generic than filter and wrapper techniques. MDI and MDA methods are examples of this technique, which will be explained in detail in the next subsection.

**4. Hybrid:** In this technique, two or more filter and wrapper methods (e.g., [7]) are combined to select a subset of the features to take advantage of each method and avoid their disadvantages. It is similar to the embedded technique but is more generic.

### B. Existing Feature Engineering Methods

We chose the embedded technique for comparison with our method among the four feature selection techniques, considering their balance between accuracy and selection time. Specifically, we delve into two embedded feature selection methods, MDI and MDA. Additionally, we introduce the PCA feature extraction technique as part of our comparison, as it is commonly used in similar studies.

By introducing the three techniques, this subsection aims to clarify the differences between our method and existing methods, which will be highlighted and compared in the results section.

#### 1. MDI method:

MDI, also called Gini importance, is based on RF and is used to calculate the importance of each feature based on the weighted sum of the actual decrease in impurity for each feature across all trees [12]. The larger the MDI score, the more important the feature. Since IDS uses a binary classification model, labeling with normal and attack, the decrease in impurity ( $I$ ) can be calculated using Eq. 1:

$$I = G_{PE} - P_{LS}G_{LS} - P_{RS}G_{RS} \quad (1)$$

Here,  $G_{PE}$  is the parent Gini (G) impurity index, as shown in Eq. 2.  $G_{LS}$  and  $G_{RS}$  are G indices for the left and right splits from the parent node in the tree, and  $P_{LS}$  and  $P_{RS}$  are the proportions for each split from their parent node (i.e.,  $P_{LS} + P_{RS} = 1$ ).

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i) \quad (2)$$

Here,  $n_c$  is the number of classes, which in our case is 2, and  $p_i$  is the ratio for the  $i^{th}$  class. G equals 0.5 if the number of samples for each class is the same and 0 if only one class is found in the dataset. However, this method is known to be biased toward high cardinality features [13].

## 2. MDA method:

MDA is also called permutation importance (Perm) and is similar to MDI as both are based on RF [13]. This method requires a validation set to calculate the importance score for each feature ( $f$ ). This score is the weighted difference between the model's prediction error rate for the validation set before and after the permutation of each feature  $f$  across all the trees, as shown in Eq. 3:

$$MDA \text{ score}_f = \frac{1}{n_t} \sum_{j=1}^{n_t} (sa_{jf} - sb_{jf}) \quad (3)$$

Here  $n_t$  is the number of trees,  $sa_{jf}$  and  $sb_{jf}$  are the scores after and before permutating feature  $f$  in the  $j^{th}$  tree, respectively. Similar to MDI, the larger the score, the more important the feature. In general, MDA can result in ignoring more irrelevant features than the MDI method.

## 3. PCA method:

PCA differs from the previous two methods since it creates new features from the original ones. These new features are called principal components (PCs) that are uncorrelated and represented by a set of eigenvectors [10]. These eigenvectors and their corresponding eigenvalues are calculated using a covariance matrix. The PCs are sorted in descending order based on their explained variance, where the first PC has the highest explained variance among all features. The explained variance for each PC ( $var\_PC_i$ ) is the ratio of that PC's eigenvalue ( $\lambda_i$ ) to the sum of all eigenvalues, as shown in Eq 4.

$$var\_PC_i = \frac{\lambda_i}{\sum_{i=1}^{n_{pc}} \lambda_i} \quad (4)$$

The easiest and most effective way to set the required number of PCs ( $n_{pc}$ ) with good performance is by setting a threshold to calculate the necessary number of PCs to get 95%-99% explained variance [14]. PCA improves model performance and is versatile to most ML models, but it is laborious to tune the threshold.

## III. RELATED WORK

The classification of IDS can be divided into five categories: network IDS (NIDS), host IDS (HIDS), protocol-based IDS (PIDS), application protocol-based IDS (APIDS), and hybrid IDS [15]. NIDS [16], [17] are designed to monitor the network

traffic of all network communications and are usually centralized in one point of the system, such as the cloud. On the other hand, HIDS [18] only monitors the traffic of only one device. PIDS [19], [20] and APIDS are set up to monitor specific protocol connections, e.g., hypertext transfer protocol secure (HTTPS), and application-specific protocols, e.g., structured query language (SQL), respectively. Hybrid IDS integrates multiple IDS mentioned above to leverage each IDS type's strengths.

Most of the IDS in the IoMT systems are NIDS since the extensive infrastructure of IoMT systems requires IDS that can monitor the whole network. *While various types of IDS exist, our method generically applies to all of them. To prevent any confusion, it is important to clarify that the main focus of this paper is on feature engineering techniques for IDS in IoT systems. The intention is not to introduce a new IDS for IoT but to propose a generic feature engineering approach that can be applied effectively in IoT environments.*

Different prior works have shown the importance of feature engineering in improving the IDS's performance [21] in the context of IIoT security, such as supervisory control and data acquisition (SCADA) systems [22] and cloud security [23]. According to Hakim et al. [21], feature engineering has improved some of the tested models' accuracies from 51% to 97%. Also, the required training time in all models has been almost reduced by half. Thus, developing a feature reduction or feature extraction approach to enhance ML models' performance is commonly recommended [24].

Improving IDS's performance can be achieved by using one or more feature reduction methods (i.e., the hybrid technique introduced in Subsection II.A), as discussed in [8], [25]–[27]. Ravindranath et al. [25] propose a feature reduction method that utilizes the whale Pearson hybrid wrapper. This method is based on the binary Whale optimization algorithm, a swarm intelligence algorithm. It reduces the data features from 42 features in the HackerEarth network attack prediction dataset [28] to only 8, with an 8% accuracy improvement compared to the original dataset using the k-nearest neighbors algorithm. Padmashree and Krishnamoorthi [26] propose a decision tree-based Pearson correlation recursive feature elimination model to select a subset of the features to detect various attacks via an optimized deep neural network (DNN) model using the BoT-IoT dataset. This model reduces the number of features in the BoT-IoT dataset to only nine features with 99.20% accuracy.

Kamarudin et al. [8] combine filter and wrapper methods as a single hybrid feature reduction method. This method reduced the number of features from 41 and 33 to 12 and 5 for the KDD CUP'99 and DARPA 1999 datasets, respectively. Also, it enhanced the IDS performance by 9%.

Another feature reduction method for IDS developed by Pawar et al. [27] selects a subset of features based on a voting scheme from a list of feature selection methods. This scheme reduced the number of features from 41 to 14 for the NSL-KDD dataset and 47 to 18 for the UNSW-NB15 dataset. Yet, none of these feature engineering methods are designed to work on IoT systems.

Another way to design a hybrid feature reduction method is by combining a feature reduction method with some specific

artificial neural network (ANN) algorithm. Jingyi et al. [29] implement a method based on supervised locality-preserving projections and use a backpropagation neural network called an extreme learning machine. Madanan et al. [30], Abdul Lateef et al. [31], Fatani et al. [32], and Dahou et al. [33] also design similar methods using intelligent water drops, crow swarm optimization algorithms, Aquila optimizer, and reptile search algorithm, respectively. While these methods use the KDD CUP'99 dataset, the Fatani et al. work included three other datasets, including NSL-KDD, BoT-IoT, and CIC2017. Using the KDD CUP'99 dataset, the Fatani et al. method performed the best with an accuracy score of 99.92% compared to 99.56%, 92.34%, 98.58%, and 98.34% for Madanan et al., Dahou et al., Jingyi et al., and Abdul Lateef et al., respectively. However, these methods must work with ANN models, which require significant computing power and only work on powerful devices.

Hybrid feature reduction methods are also used to improve the detection rate for medical diagnostics, such as [7], [9]. Shaban et al. [7], similar to [8], employ filter and wrapper methods to improve the performance of a KNN model, which is used as a new COVID-19 detection strategy. On the other hand, Li et al. [9] illustrate that using multiple feature reduction methods, including PCA in a support vector machine model, can enhance the detection rate for sleep apnea. Nimbalkar et al. [34] propose a hybrid feature selection method based on the information gain and gain ratio methods to detect DoS and DDoS attacks in IoT systems using the BoT-IoT and KDD Cup 1999 datasets.

In general, our method stands out from other approaches as it significantly improves the performance of IDS in IoT systems. Also, It takes the benefits of both feature selection and extraction methods and reduces their drawbacks. LEMDA is based on embedded methods and achieves a better tradeoff between performance and speed. It offers support for various types of attacks without needing a specific ML model or a complex ANN model. Additionally, it often leads to faster IDS models compared to alternative methods.

#### IV. OUR PROPOSED METHOD

Designing a new feature reduction method is essential to enhance the prediction for IDS, especially for IoT systems, since they require fast detection. This makes high accuracy and fast execution indispensable for IDS models.

Our method, LEMDA, is based on MDA and consists of two techniques to satisfy the high accuracy and fast speed requirements of IoT-oriented IDS. The main technique is WEDF, which runs after MDA, where the list of the most informative features is selected. The second one, SF, is an add-on technique to handle the datasets with a categorical feature as the most important feature for the cases when there are a majority of passive attacks like sniffing. In this section, we explain in detail these two techniques. For the rest of the paper, we will use  $f_m$  to represent the most informative feature in the list, which is selected by the MDA method, and  $f_{mn}$  to represent a new feature, which is created by the WEDF method.

##### A. Weighted Exponential Decay Formula (WEDF)

WEDF creates a new feature  $f_{mn}$  based on a predefined dictionary (*WEDF*). This dictionary is constructed from  $f_m$  by transforming its samples' values into weights using the exponential decay formula (Eq. 5).

$$f(x) = ab^x \quad (5)$$

Here,  $f(x)$  is the output value (after the decay) in the exponential decay formula,  $a$  is the initial value (before the decay),  $b$  is the decay factor (a static fraction,  $0 < b < 1$ , that needs to be set before running the WEDF method e.g.,  $b = 0.5$ ), and  $x$  is the time period (during which  $a$  has been decayed). Since  $a$  is a static parameter, it can be removed in the WEDF method (i.e., considering  $a = 1$ ).

$$WEDF_u = f(p)w_u = b^p w_u \quad \text{where} \quad w_u = \frac{z_u}{n_u} \quad (6)$$

Eq. 6 calculates  $WEDF_u$ , the WEDF score for each  $u$  that constitutes *WEDF*.  $u$  is a specific unique value from all data instances of the  $f_m$  feature. Each  $u$  corresponds to a unique data value. In the context of the WUSTL-EHMS dataset, for instance,  $u$  can be "TCP," which is a value of the  $f_m$  feature. A more detailed example is provided in the next paragraph. We add a new weight parameter  $w_u$  for each unique value  $u$  in  $f_m$ .  $z_u$  represents the number of attack samples in the training dataset for each  $u$  in  $f_m$ , and  $n_u$  is the total number of samples for each  $u$  in  $f_m$ . Hence,  $z_u$  divided by  $n_u$  will result in  $w_u$  for each  $u$ . All the weights are sorted in descending order based on  $n_u$ . Let us denote the index of each  $u$  as  $p$  (i.e.,  $p$  ranges from 1 to the number of unique values in  $f_m$ ).

For instance, let us assume that 100 out of 1000 samples in the training dataset have transmission control protocol (TCP)  $u = TCP$  as their unique value, 10 of which are attack samples; then  $z_{TCP} = 10$ ,  $n_{TCP} = 100$ , and  $w_{TCP} = 10/100 = 0.1$ , which will be stored in the  $w$  dictionary. Then, assuming TCP is the first unique value in the  $w$  (i.e.,  $p = 1$ ) and setting  $b = 0.5$ , we can calculate  $WEDF_{TCP}$  by  $b^p w_{TCP} = 0.5^1 \times 0.1 = 0.05$ . Hence, all the samples with  $u = TCP$  will have  $WEDF_{TCP} = 0.05$ . Other unique values in  $f_m$  with zero attack samples will have a WEDF score of zero in  $f_{mn}$ .

Finally, the WEDF scores for the  $u$  values in the  $f_m$  feature using the training dataset are stored in a dictionary (*WEDF*). This dictionary is used to create the  $f_{mn}$  feature in the training and testing datasets. Then, the  $f_m$  feature is deleted from both datasets. Algorithm 1 shows the step-by-step implementation of generating the dictionary.

Upon generating the dictionary (assigning a weight to each  $u$  based on the proportion of attack samples associated with each value, i.e.,  $w_u$  to attack samples and 0 to normal samples), the feature distribution for the  $f_{mn}$  feature will become roughly a bimodal distribution. Consider an example of a  $f_m$  with a standard normal distribution  $N(0, 1)$ . After applying WEDF, the normal samples will tend to cluster around the left side (0.1% region) of the distribution, and the attack samples will cluster around the right side of the distribution, resulting in a gap in the distribution between the normal and

attack samples. This helps the ML model to easily separate the normal samples from the attack samples and increases the importance of the  $f_{mn}$  feature compared to the original  $f_m$  feature, resulting in better IDS performance.

---

**Algorithm 1** WEDF Method
 

---

```

1: Input:  $f_m$  feature from the Training Dataset ( $D_T$ )
2: Output:  $WEDF$  for all  $WEDF_u$  in  $f_m$  to create  $f_{mn}$ 
3:  $WEDF, z, w, w_{attack}, w_{normal} = \{\}, \{\}, \{\}, \{\}, \{\}$ 
4:  $n \leftarrow value\_counts(D_T[f_m])$ 
5: for  $u$  in  $n$  do
6:    $z[u] \leftarrow length(D_T[f_m][u], label = attack)$ 
7:   if  $z_u \geq 1$  then
8:      $w_{attack}[u] \leftarrow z[u]/n[u]$ 
9:   else
10:     $w_{normal}[u] \leftarrow 0$ 
11:   end if
12: end for
13:  $w \leftarrow concatenate(w_{attack}, w_{normal})$ 
14:  $p = 1$ 
15: for  $u$  in  $w$  do
16:    $WEDF[u] \leftarrow b^p \times w[u]$ 
17:    $p = p + 1$ 
18: end for

```

---

### B. Sensitivity Factor (SF)

SF has been added as an add-on besides using the WEDF technique in case the  $f_m$  feature is a categorial feature like the *Flags* feature in networking, and most attacks are passive attacks like sniffing. This add-on requires the training and testing datasets, individually, to be arranged in a sequential order, typically based on the timestamps associated with the samples.

SF is also based on the exponential decay formula but without multiplying weights  $w$ . As shown in Eq. 7, we use  $d$ , an index of the current sample ( $s$ ) based on the last seen suspicious sample, as the input to the exponential decay formula (Eq. 5). Similar to the WEDF method, a new feature  $f_{smn}$  is created in the training and testing datasets using the  $f_m$  feature. Using the *Flags* feature in networking as an example, any item in *Flags* that differs from the common values of *Flags*, such as duplicate MACs (M), is considered a suspicious sample. This add-on is used in one of the three datasets, and its results are promising, as shown in the results section.

$$SF_s = b^d \quad (7)$$

In the case of suspicious samples, the SF score reaches its peak (i.e., a higher SF score indicates a greater likelihood of being an attack sample). Then, the score exponentially decreases for each sample after the suspicious sample(s) until the score reaches zero, as shown in Algorithm 2. This is because cyber attacks usually exhibit intensive behaviors over a continuous time period, and the network traffic returns to a normal state after a certain duration.

---

**Algorithm 2** SF Method
 

---

```

1: Input:  $f_m$  feature from the Training Dataset ( $D_T$ )
2: Output:  $f_{smn}$  feature
3:  $common\_value \leftarrow$  the most common value in the normal samples
4:  $b, d = 0, 1$ 
5: for  $s$  in  $f_m$  do
6:   if  $s$  is not a  $common\_value$  in  $f_m$  then
7:      $b, d = 0.5, 1$ 
8:   end if
9:    $s_{new} \leftarrow b^d$ 
10:   $d = d + 1$ 
11: end for

```

---

## V. EXPERIMENTAL METHODOLOGY

In this section, we will demonstrate the evaluation results of the proposed method using three datasets and three models. The datasets are WUSTL-EHMS [11], MQTT-IoT [35], and BOT-IoT [36]. Two of them are collected from general IoT systems (MQTT-IOT and BOT-IoT) and one from an IoMT system (WUSTL-EHMS). Our ML models include decision trees (DT), RF, and two ANN models. We use the  $F_1$  score, the Safety score [43], and the accuracy metrics to compare the performance of feature engineering methods applied to these models.

Using these datasets and models, we evaluate our methods by comparing them to two widely recognized feature reduction techniques, namely PCA and MDA, as well as the scenario where no feature reduction (Base) is applied.

### A. Datasets

The three IoT datasets used in our experiments have different sizes starting from 16K to 10M samples and similar numbers of features, as shown in Table I:

TABLE I  
DATASETS STATISTICS

Dataset	No. of Samples	No. of Features
WUSTL-EHMS	16,317	44
MQTT-IoT	2,000,000	31
BOT-IoT	10,000,000	35

#### 1. WUSTL-EHMS:

This dataset was collected from a real-time enhanced healthcare monitoring system (EHMS) testbed at Washington University in St. Louis, representing a real IDS for the IoMT systems [11]. The types of attacks in this dataset are based on man-in-the-middle (MiTM) attacks, such as sniffing and injection attacks. Hence, this dataset has both passive (sniffing only) and active (injection) attacks. This dataset is explained in [11] and [37].

#### 2. MQTT-IoT:

This dataset uses message queuing telemetry transport (MQTT) protocol for IoT systems [35]. The types of attacks in this dataset are as follows: user datagram protocol (UDP) scan, aggressive scan, MQTT brute-force Sparta, and secure shell

protocol (SSH) brute-force. The number of samples in this dataset is 20M, but we have randomly selected 2M samples containing all attacks in the original dataset. More about this dataset is available in [35], [38].

**3. BOT-IoT:**

This well-known dataset was created using IDS for IoT systems in the Cyber Range Lab of UNSW Canberra. It has different types of attacks, including theft, reconnaissance, denial of service (DoS), and distributed DoS (DDoS) attacks. 10M samples out of 73M samples have been selected to test our method. More about this dataset is available in [36], [39].

The number of selected features for each feature reduction method using these datasets is presented in Table II. The Base method represents the method where we use all the features (without feature engineering). It is worth noting that we have removed the identifier features, such as IP addresses and port numbers, from all the methods, including the Base method.

TABLE II  
NUMBER OF SELECTED FEATURES PER REDUCTION METHOD

Dataset	Base	PCA*	MDA	LEMMA
WUSTL-EHMS	35	14	5	5+1**
MQTT-IoT	25	9	5	5
BOT-IoT	23	10	5	5

\* Explained variance = 95%

\*\* Additional feature using the SF method

**B. Models**

We have used DT, RF, and two ANN models. The scikit-learn package has been utilized for DT and RF models with default hyperparameters [40] [41]. A simple multi-layer perceptron (MLP) model with two layers is used as a simple ANN model. To show the effect of complex ANN models, we have added a Convolutional Neural Network (CNN) model with five layers and Max pooling layers. The Keras package has been utilized for the two ANN models [42]. More about the hyperparameters are shown in Table III.

As our objective is not to develop optimal machine learning models with optimized hyperparameters but to demonstrate our method’s robustness and maintain consistency in experimental comparisons with other methods, we use these simple models with identical parameters across all three datasets.

TABLE III  
ANN MODELS HYPERPARAMETERS

Parameter	MLP Typical Value(s)	CNN Typical Value(s)
# of layers	2	5
# of neurons per layer	20, 1	32, 128, 512, 1024, 1
# of epochs	20	2
Kernel size	None	2
Pooling	None	Max Pooling (size=2)
Batch size		1000
Loss function		binary cross-entropy
Optimizer		Adam
Activation function		tanh, sigmoid

The k-fold cross-validation method with ten folds has been utilized on all three models to analyze the models’ performance. In each fold, the dataset is divided into ten subsets; nine of them are used for training and one for testing.

**C. Metrics**

We evaluate the models’ performances using the accuracy,  $F_1$  score, and Safety score. All the metrics are calculated based on the following four categories: true negative (TN), true positive (TP), false positive (FP), and false negative (FN). Label attack is defined as positive, and label normal is defined as negative. TN and TP are the cases when an IDS model correctly predicts a normal sample as normal and an attack sample as an attack, respectively. FP is when the model mistakenly predicts a normal sample as an attack, while FN is when an attack sample is predicted as normal.

**1. Accuracy:**

This metric represents the ratio of correct predictions to the total predictions, as illustrated in Eq. 8.

$$Accuracy = \frac{TN + TP}{FP + FN + TN + TP} \tag{8}$$

**2.  $F_1$  score:**

This metric is popular for security applications such as IDS. It is the harmonic mean between recall and precision, as illustrated in Eq. 9.

$$F_1 \text{ score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP+FN)} \tag{9}$$

**3. Safety score:**

The safety score is designed by Salman et al. [43] specifically for security applications to fulfill the shortcoming of other metrics in these types of applications. This metric adds weights (Eq. 10) to the primary four model’s outcome categories (TP, TN, FP, and FN).

$$Safety \text{ score} = \frac{w_{TN}TN + w_{TP}TP}{w_{FP}FP + w_{FN}FN + w_{TN}TN + w_{TP}TP} \tag{10}$$

For generality, we assign the following weights assuming that both FN and FP have the same importance, as explained in [11] and [43]:

$$w_{TN} = \frac{1}{19}, w_{TP} = \frac{2}{19}, w_{FP} = \frac{8}{19}, w_{FN} = \frac{8}{19}$$

TABLE IV  
LIST OF SELECTED FEATURES

#	Dataset		
	WUSTL-EHMS	MQTT-IoT	BOT-IoT
1	<i>Flgs</i>	<i>protocol</i>	<i>state</i>
2	<i>DIntPkt</i>	<i>mqtt_messageType</i>	<i>sbytes</i>
3	<i>DstJitter</i>	<i>ttl</i>	<i>bytes</i>
4	<i>Rate</i>	<i>mqtt_messageLength</i>	<i>proto</i>
5	<i>DstLoad</i>	<i>ip_len</i>	<i>srate</i>



versa.

The selected features using the MDA method and our method for all three datasets are shown in Table IV. Note that our method creates an  $f_{mn}$  feature for each dataset using their  $f_m$  features, which are shown in the first row of Table IV.

#### A. WUSTL-EHMS:

The WUSTL-EHMS dataset, as shown in Table IV, has the  $Flgs$  feature as the most informative feature  $f_m$ , and the majority of the attacks are passive attacks (sniffing). As a result, among the three datasets, it is the only one that is suitable to use the SF add-on with the WEDF method. Across the three models in Table V, our method shows average values for accuracy,  $F_1$  score, and Safety score of approximately 95%, 78%, and 73%, respectively. These outcomes show that our method has an average improvement of about 28%, 52%, and 61% in accuracy,  $F_1$  score, and Safety score, respectively, compared to the other methods.

DT, RF, and ANN models show similar performances across the Base, PCA, and MDA methods, while our method outperforms all of them. As seen in Table V, our method performs almost twice better with the security-oriented metric Safety score than the other methods in the three models. With the accuracy and the  $F_1$  score, our method still significantly shows improved results compared to other methods.

#### B. MQTT-IoT:

Similar to the WUSTL-EHMS dataset results, across the three models, our method showcases average accuracy,  $F_1$  score, and Safety score of approximately 91%, 95%, and 73% across the three models. MQTT-IoT results showed that our method has outperformed other methods by at least 50% using the  $F_1$  score, as illustrated in Table V. Furthermore, even when the MDA method uses the same 4 out of 5 features as our method, the difference in performance between them reached almost 70% on average. The average improvements of our method using the accuracy and Safety score are 56% and 79%, respectively.

#### C. BOT-IoT:

This dataset has more attacks, such as DoS and DDoS attacks, making it more general for IoT systems than the other two datasets. Our method demonstrates an average performance of approximately 99% for accuracy, 99% for the  $F_1$  score, and 98% for the Safety score across the three models. Even here, our method shows clear performance improvement in the DT and RF models compared to the other methods in terms of accuracy,  $F_1$  score, and Safety score, as shown in Table V.

Given these results and the varying attacks in each dataset, our method demonstrates superior performance, rendering it more suitable as an IDS for IoT systems using AI-based models compared to other methods.

In particular, by comparing the results of the MLP and CNN models across all three datasets, the CNN exhibits

superior performance over the MLP in the two large datasets, MQTT-IoT and Bot-IoT. This suggests that for large datasets without feature engineering, a more complex ANN model is beneficial. Additionally, LEMDA exhibits enhancement (from MLP to CNN) in two datasets, WUSTL-EHMS and MQTT-IoT. However, this increase in complexity leads to very high training and detection times.

These findings confirm that feature engineering methods are essential to reduce the computational complexity with simpler models. While more complex models may not necessarily enhance performance, they still contribute to this reduction of computing time.

#### D. Training and Detection Time Comparison:

The improvement in model performance is not the only important requirement for IDS in IoT systems since the models' training and detection time are also critical. As presented in Table V, our method achieves the lowest or very close to the lowest training time compared to other methods, with an average of 0.66s, 34.04s, and 199.76s in WUSTL-EHMS, MQTT-IoT, and BOT-IoT datasets, respectively.

The detection times are also shown in Table V. Similar to the average training time, our model detection times are the lowest in almost all the cases, with an average of 0.05s, 1.62s, and 8.10s in WUSTL-EHMS, MQTT-IoT, and BOT-IoT datasets, respectively. This lets us conclude that our method enhances the IDS performance and takes less time to train and detect attacks using different models in most cases. Hence, it makes the IDS models very accurate and fast to train ML models and detect attacks.

#### E. Related Work Comparison:

In addition to comparing our work with PCA and MDA, we further assess its performance against four related works [26], [32], [33], and [34], using the Bot-IoT dataset, as presented in Table VI. As mentioned in Section III, [32], [33], and [34] used only one ML classifier model to report their results, while [26] method uses DNN classifier. In contrast, we have tested our method with multiple ML models, including DT and RF, in addition to the two ANN models, including MLP and CNN, to show the versatility of our method.

All the attacks were included in [26], [32], and [33] results, while only DoS/DDoS attacks were included in [34]. The methods proposed by [26], [32], and [33] require using an ANN model as part of the implementation of their methods. The methods in [32] and [33] were built to transform the feature space before selecting the best set of features. On the other hand, our method will only transform the most informative feature after the selection process is completed by the MDA method. Compared to these methods, our work shows comparable or better results than these works with up to 85% feature reduction rate using ML and ANN models.

## VII. CONCLUSION

IDS models for IoT systems require faster training and detecting time along with high performance. Therefore, these




TABLE VI  
RELATED WORK COMPARISON USING BOT-IOT DATASET

	LEMMA	[26]	[32]	[33]	[34]
Category	Supervised	Supervised	Supervised	Supervised	Supervised
Type of attacks	All	All	All	All	DoS/DDoS
Model	DT/RF/ANN	DNN	KNN	KNN	JRip
Require ANN	No	Yes	Yes	Yes	No
Reduction Approach	Selection+Extraction	Only selection	Selection+Extraction	Selection+Extraction	Only selection
Reduction Method	MDA+WEDF	DT-PCRFE	AQU+CNN	RSA+CNN	IG+GR
No. of Samples	10M	13.9M	3.6M	3.6M	0.7M
Feature Reduction Rate	85.7%	74.3%	—	—	54.3%
No. of Features	5	9	—	—	16
Accuracy	99.998% [DT/RF]	99.200%	99.994%	99.993%	99.999%
$F_1$	99.999% [DT/RF]	98.910%	99.992%	99.992%	—

require specialized feature reduction methods. This paper presented a new feature reduction method called LEMMA. Our proposed method uses two new techniques called WEDF and SF to generate a representative feature based on the most informative feature from the MDA method. We used three different datasets with different sizes, three different ML models, and three different metrics. We compare our method with other methods, including MDA, PCA, and a base method without feature reduction methods as the ground truth of our experimental results. Our results show that LEMMA performs better than the other methods in all three datasets and ML models by an average of 34%, 57%, and 56% using the  $F_1$ , the Safety scores, and the accuracy scores, respectively. Furthermore, the proposed method achieved the lowest required training and detection times in most cases, making it run faster than other methods.

For future work, we plan to investigate the improvement of our method using best-optimized models and then compare the results with the plain models. We will examine the potential of our method for semi-supervised and unsupervised ML-based IDS. Moreover, applying our method to applications (other than IoT systems) can help determine its limits.

ACKNOWLEDGMENTS

s work was supported in part by NSF under Grant CNS-1718929; in part by the National Priorities Research Program (NPRP) from the Qatar National Research Fund (QNRF) under Award NPRP13S-0205-200265 (a member of the Qatar Foundation); and in part by Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia.

REFERENCES

[1] E. Schiller, A. Aidoo, J. Fuhrer, J. Stahl, M. Ziörjen, and B. Stiller, “Landscape of IoT security,” *Computer Science Review*, vol. 44, 100467, 2022.

[2] F. Li, R. Xie, Z. Wang, L. Guo, J. Ye, P. Ma, and W. Song, “Online distributed IoT security monitoring with Multidimensional Streaming Big Data,” *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4387–4394, 2020.

[3] A. Thakkar and R. Lohiya, “A review on machine learning and Deep Learning Perspectives of IDS for IoT: Recent updates, security issues, and challenges,” *Archives of Computational Methods in Engineering*, vol. 28, no. 4, pp. 3211–3243, 2020.

[4] R. Vijayanand and D. Devaraj, “A Novel Feature Selection Method Using Whale Optimization Algorithm and Genetic Operators for Intrusion Detection System in Wireless Mesh Network,” in *IEEE Access*, vol. 8, pp. 56847–56854, 2020.

[5] J. Cui, L. Zong, J. Xie, and M. Tang, “A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data,” *Applied Intelligence*, vol. 53, no. 1, pp. 272–288, 2023/01/01 2023, doi: 10.1007/s10489-022-03361-2.

[6] P. Yang, H. Huang, and C. Liu, “Feature selection revisited in the single-cell era,” *Genome Biology*, vol. 22, no. 1, p. 321, 2021.

[7] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsooud, “A new COVID-19 patients detection strategy (CPDS) based on hybrid feature selection and Enhanced KNN classifier,” *Knowledge-Based Systems*, vol. 205, p. 106270, 2020.

[8] M. H. Kamarudin, C. Maple, and T. Watson, “Hybrid feature selection technique for intrusion detection system,” *International Journal of High-Performance Computing and Networking*, vol. 13, no. 2, pp. 232–240, 2019.

[9] X. Li, S. H. Ling, and S. Su, “A Hybrid Feature Selection and Extraction Methods for Sleep Apnea Detection Using Bio-Signals,” *Sensors*, vol. 20, no. 15, p. 4323, Aug. 2020.

[10] S. S. Shekhawat, H. Sharma, S. Kumar, A. Nayyar and B. Qureshi, “bSSA: Binary Salp Swarm Algorithm With Hybrid Data Transformation for Feature Selection,” in *IEEE Access*, vol. 9, pp. 14867–14882, 2021.

[11] A. A. Hady, A. Ghubaish, T. Salman, D. Unal, and R. Jain, “Intrusion Detection System for Healthcare Systems Using Medical and Network Data: A Comparison Study,” in *IEEE Access*, vol. 8, pp. 106576–106584, 2020.

[12] Y. A. Farrukh, Z. Ahmad, I. Khan, and R. M. Elavarasan, “A Sequential Supervised Machine Learning Approach for Cyber Attack Detection in a Smart Grid System,” 2021 North American Power Symposium (NAPS), 2021, pp. 1–6.

[13] I. R. Ward, L. Wang, J. Lu, M. Bennamoun, G. Dwivedi, and F. M. Sanfilippo, “Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes?” *Computer Methods and Programs in Biomedicine*, vol. 212, 2021, p. 106415.

[14] S. B. Vilsen and D.-I. Stroe, “Battery state-of-health modelling by multiple linear regression,” *Journal of Cleaner Production*, vol. 290, p. 125700, 2021.

[15] D. Swain, N. Chillur, S. Patel and A. Bhilare, “Intelligent System for Detecting Intrusion with Feature Bagging,” 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), 2021, pp. 1–4, doi: 10.1109/AIMV53313.2021.9670940.

[16] X.-H. Nguyen, X.-D. Nguyen, H.-H. Huynh, and K.-H. Le, “Real-guard: A Lightweight Network Intrusion Detection System for IoT Gateways,” *Sensors*, vol. 22, no. 2, p. 432, 2022. [Online]. Available: <https://www.mdpi.com/1424-8220/22/2/432>.

[17] J. Yu, X. Ye, and H. Li, “A high precision intrusion detection system for network security communication based on multi-scale convolutional neural network,” *Future Generation Computer Systems*, vol. 129, pp. 399–406, 2022/04/01/ 2022, doi: <https://doi.org/10.1016/j.future.2021.10.018>.

[18] D. Lightbody, D. -M. Ngo, A. Temko, C. Murphy and E. Popovici, “Host-Based Intrusion Detection System for IoT using Convolutional Neural Networks,” 2022 33rd Irish Signals and Systems Conference (ISSC), Cork, Ireland, 2022, pp. 1–7, doi: 10.1109/ISSC55427.2022.9826188.

[19] S. Mandal, A. Sai Sabitha, and D. Mehrotra, “Analysis on Protocol-Based Intrusion Detection System Using Artificial Intelligence,” in *Machine Intelligence and Smart Systems*, Singapore, S. Agrawal, K. Kumar Gupta, J. H. Chan, J. Agrawal, and M. Gupta, Eds., 2021// 2021: Springer Nature Singapore, pp. 131–143.

- [20] M. Zeeshan et al., "Protocol-Based Deep Intrusion Detection for DoS and DDoS Attacks Using UNSW-NB15 and Bot-IoT Data-Sets," in *IEEE Access*, vol. 10, pp. 2269-2283, 2022, doi: 10.1109/ACCESS.2021.3137201.
- [21] L. Hakim, R. Fatma, and Novriandi, "Influence Analysis of Feature Selection to Network Intrusion Detection System Performance Using NSL-KDD Dataset," 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), 2019, pp. 217-220.
- [22] D. Upadhyay, J. Manero, M. Zaman, and S. Sampalli, "Gradient Boosting Feature Selection With Machine Learning Classifiers for Intrusion Detection on Power Grids," in *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 1104-1116, 2021.
- [23] P. Mishra, V. Varadharajan, E. S. Pilli, and U. Tupakula, "VMGuard: A VMI-Based Security Architecture for Intrusion Detection in Cloud Environment," in *IEEE Transactions on Cloud Computing*, vol. 8, no. 3, pp. 957-971, 1 July-Sept. 2020.
- [24] T. Parlar and E. Sarac, "IWD based feature selection algorithm for sentiment analysis," *Elektronika ir Elektrotechnika*, vol. 25, no. 1, 2019.
- [25] V. Ravindranath, S. Ramasamy, R. Somula, K. S. Sahoo, and A. H. Gandomi, "Swarm Intelligence Based Feature Selection for Intrusion and Detection System in Cloud Infrastructure," 2020 IEEE Congress on Evolutionary Computation (CEC), 2020, pp. 1-6.
- [26] A. Padmashree and M. Krishnamoorthi, "Decision Tree with Pearson Correlation-based Recursive Feature Elimination Model for Attack Detection in IoT Environment," *Information Technology and Control*, vol. 51, no. 4, pp. 771-785, 2022.
- [27] Y. Pawar, N. Zamzami, and N. Bouguila, "An Effective Hybrid Anomaly Detection System Based on Mixture Models," 2020 International Symposium on Networks, Computers and Communications (ISNCC), 2020, pp. 1-6.
- [28] HackerEarth, "Predict Network Attacks," 2020. [Online]. Available: <https://www.hackerearth.com/problem/machine-learning/sample/>. [Accessed: 03-Oct-2023].
- [29] W. Jingyi, G. Xusheng, H. Jieli, and L. Shenghou, "ELM Network Intrusion Detection Model Based on SLPP Feature Extraction," 2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2021, pp. 46-49.
- [30] M. Madanan, A. Venugopal, and N. C. Velayudhan, "A Hybrid Anomaly Based Intrusion Detection Methodology Using IWD for LSTM Classification," 2020 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 2020, pp. 1-5.
- [31] A. A. Abdul Lateef, S. T. Faraj Al-Janabi, and B. Al-Khateeb, "Hybrid Intrusion Detection System Based on Deep Learning," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-5.
- [32] A. Fatani, A. Dahou, M. A. A. Al-qaness, S. Lu, and M. A. Abd Elaziz, "Advanced Feature Extraction and Selection Approach Using Deep Learning and Aquila Optimizer for IoT Intrusion Detection System," *Sensors*, vol. 22, no. 1, p. 140, Dec. 2021, doi: 10.3390/s22010140.
- [33] A. Dahou et al., "Intrusion Detection System for IoT Based on Deep Learning and Modified Reptile Search Algorithm," *Computational Intelligence and Neuroscience*, vol. 2022, p. 6473507, 2022/06/02 2022, doi: 10.1155/2022/6473507.
- [34] P. Nimbalkar and D. Kshirsagar, "Feature selection for intrusion detection system in Internet-of-Things (IoT)," *ICT Express*, vol. 7, no. 2, 2021, pp. 177-181.
- [35] H. Hindy, E. Bayne, M. Bures, R. Atkinson, C. Tachtatzis, and X. Bellekens, "Machine learning based IOT intrusion detection system: An MQTT case study (MQTT-IOT-IDS2020 dataset)," *Selected Papers from the 12th International Networking Conference*, pp. 73-84, 2021.
- [36] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, 2019, pp. 779-796.
- [37] Raj Jain, "WUSTL EHMS 2020 Dataset for Internet of Medical Things (IoMT) Cybersecurity Research," 2020. [Online]. Available: <https://www.cse.wustl.edu/~jain/ehms/index.html>. [Accessed: 03-Oct-2023].
- [38] H. Hindy, C. Tachtatzis, R. Atkinson, E. Bayne, X. Bellekens, June 23, 2020, "MQTT-IoT-IDS2020: MQTT Internet of Things Intrusion Detection Dataset," *IEEE Dataport*.
- [39] N. Moustafa, October 16, 2019, "The Bot-IoT dataset," *IEEE Dataport*.
- [40] scikit-learn "DecisionTreeClassifier," 2023. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>. [Accessed: 03-Oct-2023].
- [41] scikit-learn, "RandomForestClassifier," 2023. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>. [Accessed: 03-Oct-2023].
- [42] Keras, "The Sequential model," 2020. [Online]. Available: [https://keras.io/guides/sequential\\_model/](https://keras.io/guides/sequential_model/). [Accessed: 03-Oct-2023].
- [43] T. Salman, A. Ghubaish, D. Unal, and R. Jain, "Safety Score as an Evaluation Metric for Machine Learning Models of Security Applications," in *IEEE Networking Letters*, vol. 2, no. 4, pp. 207-211, 2020.