A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015

Although KDD99 dataset is more than 15 years old, it is still widely used in academic research. To investigate wide usage of this dataset in Machine Learning Research (MLR) and Intrusion Detection Systems (IDS); this study reviews 149 research articles from 65 journals indexed in Science Citation In- dex Expanded and Emerging Sources Citation Index during the last six years (2010–2015). If we include papers presented in other indexes and conferences, number of studies would be tripled. The number of published studies shows that KDD99 is the most used dataset in IDS and machine learning areas, and it is the de facto dataset for these research areas. To show recent usage of KDD99 and the related sub-dataset (NSL-KDD) in IDS and MLR, the following de- scriptive statistics about the reviewed studies are given: main contribution of articles, the applied algorithms, compared classification algorithms, software toolbox usage, the size and type of the used dataset for training and test- ing, and classification output classes (binary, multi-class). In addition to these statistics, a checklist for future researchers that work in this area is provided.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1954v1 | CC-BY 4.0 Open Access | rec: 14 Apr 2016, publ: 14 Apr 2016

Noname manuscript No.
(will be inserted by the editor)

A Review of KDD99 Dataset Usage in Intrusion Detection and Machine Learning between 2010 and 2015

Atilla Özgür · Hamit Erdem

the date of receipt and acceptance should be inserted later

Abstract Although KDD99 dataset is more than 15 years old, it is still widely used in academic research. To investigate wide usage of this dataset in Machine Learning Research (MLR) and Intrusion Detection Systems (IDS); this study reviews 149 research articles from 65 journals indexed in Science Citation Index Expanded and Emerging Sources Citation Index during the last six years (2010–2015). If we include papers presented in other indexes and conferences, number of studies would be tripled. The number of published studies shows that KDD99 is the most used dataset in IDS and machine learning areas, and it is the de facto dataset for these research areas. To show recent usage of KDD99 and the related sub-dataset (NSL-KDD) in IDS and MLR, the following descriptive statistics about the reviewed studies are given: main contribution of articles, the applied algorithms, compared classification algorithms, software toolbox usage, the size and type of the used dataset for training and testing, and classification output classes (binary, multi-class). In addition to these statistics, a checklist for future researchers that work in this area is provided.

Keywords Machine Learning · KDD99 · Review · Intrusion Detection · Supervised Learning · Classification

1 Introduction

Internet, mobile, e-commerce, PC based communication, and information systems have become parts of daily life. Wide usage of these systems makes communication easier, increases data transfer and information sharing, and improves life quality. Although these systems are used in many fields, they

Atilla Özgür
Başkent University
E-mail: aozgur@baskent.edu.tr

Hamit Erdem
Başkent University E-mail: herdem@baskent.edu.tr

suffer from the various attacks such as viruses, worms, Trojan horses. Due to importance of these systems, these attacks must be identified and stopped as soon as possible. Research about finding attacks and removing their effects have been defined as Intrusion Detection Systems (IDS).

IDS studies can be considered as a classification task that separates normal behavior of networks from attacks. After the first paper about IDS [7], hundreds of studies have been published in this domain. Among other techniques, machine learning and data mining algorithms are widely used in IDS. Most of these algorithms are based on the assumption that problem space does not change very fast. But in IDS domain, attackers continuously change and improve their capabilities [22]. Due to this reason, even though machine learning and data mining algorithms are very successful in other domains, their performance decreases in IDS. Thus, IDS is an unsolved problem since this domain is an evolving problem [22].

Similar to other classification and clustering problems, IDS algorithms need training dataset to properly function. Although common and standard datasets are available for other fields, there is no recent and common dataset for IDS. Lack of a common and recent dataset for IDS research, has been mentioned by numerous studies [1, 22, 3]. Recent reviews [5, 10] also identify this problem as a research gap. Therefore, KDD99 is the most used dataset in IDS domain [9, 14, 23].

KDD99 dataset, created in 1999, is very old for IDS studies[3]. Nonetheless, it has been used in many studies during last 16 years, and cited in many studies —Reference article for KDD99 preparation [15] has been cited 873 times according to Google Scholar (February 2016). Moreover, 149 research articles that used KDD99 were published in Science Citation Index Expanded (SCIE) and Emerging Sources Citation Index (ESCI) journals between 2010 and 2015, Figure 1. Figure 1 shows that KDD99 has been frequently used in IDS or similar studies. According to results, KDD99 dataset is primarily used in IDS and machine learning research. Additionally, this dataset also has been used in other domains, such as feature selection and data streams. Regarding to Figure 2, based on the 149 published studies, 142 of them has been applied in either in IDS or in machine learning, and 118 indexed articles use two domains in the same study. These numbers shows that the main intersection of machine learning research, IDS, and information security is KDD99 dataset.

As presented in the study [2], investigating one of the most used datasets and considering applied approaches can be a subject of a research in the applied area. In a similar way, although KDD99 has been used in many IDS and machine learning studies, there is not a review study to evaluate and analyze the published research and answer the following questions:

- Which machine learning algorithms and IDS methods are used mostly?
- What is the training and testing dataset usage in the published studies?
- What are the sizes of training and testing dataset in proposed studies?
- How many classes have been considered in IDS classification?

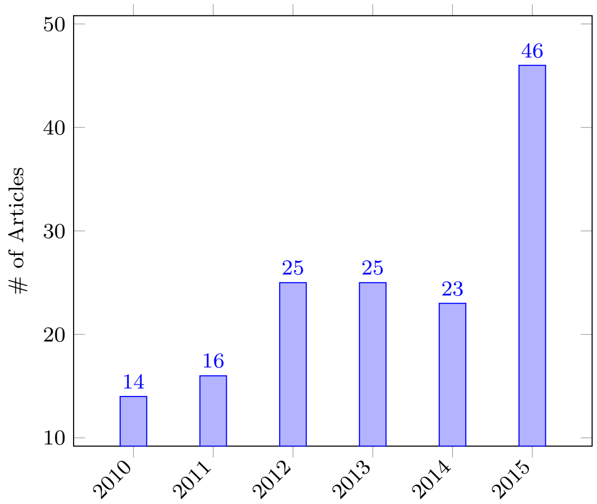

Fig. 1 KDD99 Dataset Usage By Years.

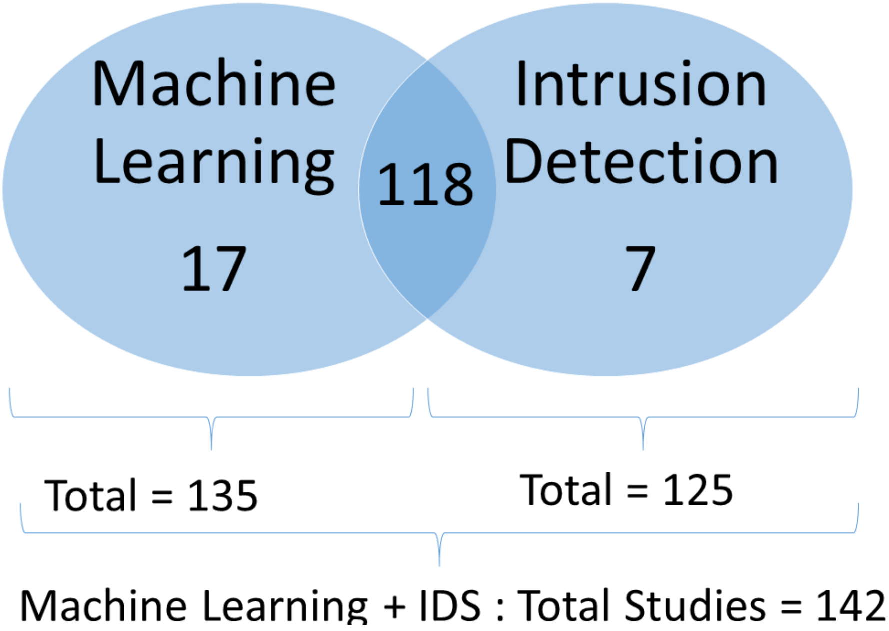

Fig. 2 Machine Learning and IDS Usage in this Review. Total of 142 articles use KKD99 in either Machine Learning or IDS between 2010 and 2015 from 149 research articles.

– Which performance metrics have been used to measure the results of the classification?
– Which software tools have been used for implementation and comparison?

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1954v1 | CC-BY 4.0 Open Access | rec: 14 Apr 2016, publ: 14 Apr 2016

To answer these questions, the proposed study reviews 149 studies from 2010 to 2015 focusing on KDD99 usage statistics. The authors think that the results of the proposed study will be useful for the other researchers who plan to use this dataset in IDS or machine learning studies.

This review differs from the previous review articles considering following aspects: First, most of the reviews in this domain try to include critical papers and explain major approaches. In contrast, our study tries to be comprehensive. Second, only articles from 65 journals (SCIE and ESCI) are included, Table 11. No conference articles are included in this study. We believe that our study includes almost all of SCI-indexed studies that used KDD99 between 2010–2015. Third, comprehensive descriptive statistics about KDD99, machine learning and IDS are given. Some of these statistics are as follows:

1. KDD99 has been analyzed considering number of output classes, training and testing datasets in reviewed studies, Table 4 and Table 5.
2. Main contribution that concerns on the applied method using KDD99. The applied methods may be clustering, classification, feature selection/reduction algorithms. All the applied methods in the focused period has been evaluated and presented in Table 3.
3. The usage frequency of machine learning and IDS algorithms has been presented in Table 6 and discussed in detail.
4. Proposed algorithms are implemented and compared with standard algorithms using variety of software packages (Table 8).
5. Training and testing dataset sizes and classification types (binary, multiclass).
6. Most of the reviewed articles compare their proposed method with other classifiers. These classifiers have been shown in Table 7 and discussed.
7. Although, KDD99 and related sub-set have been used in recent studies, some studies compared their results with other datasets, Table 9.
8. Categorizing the main theme of the published article in three main groups as Machine Learning, Anomaly Detection or Alert Correlation has been presented in Figure 7.

Finally, considering collected statistics, strengths and weaknesses of reviewed articles, a checklist is provided.

The findings of this review would be useful for researchers who may want to use KDD99 or a similar big dataset in their research since KDD99 is one of the biggest datasets in UCI repository.

The remainder of the paper is organized as follows: Section 2 considers similar related reviews. Section 3 gives definitions and history of DARPA, KDD99 and NSL-KDD datasets. Section 4 gives a general machine learning model while using KDD99, and evaluates contribution of reviewed articles considering the structure of the presented model. Section 5 gives descriptive statistics about general KDD99 usage with figures and tables. Section 6 suggests a checklist considering common mistakes and strengths points of the reviewed articles for further studies in order to improve the quality of similar studies. Finally, section 7 discusses the results of this review.

2 Related Reviews About KDD99 and Intrusion Detection

Most of the IDS reviews try to find prominent papers about the subject and summarize them. This approach provides fast learning opportunity for the reader. In contrast to previous review studies, this study follows a different approach, and tries to provide descriptive statistics for who want to use KDD99 in their research.

One of the most similar review to ours has been presented by [24] in Expert Systems with Applications in 2009. Their study evaluated 55 articles between 2000 and 2007 that focused on intersection of IDS and machine learning. First, they give definition of the most used single classifiers in machine learning for IDS containing k-nearest neighbor, support vector machines, artificial neural networks, self-organizing maps, decision trees, naive bayes, genetic algorithms, fuzzy logic, hybrid classifiers, and ensemble classifiers. Second, they provide yearly statistics for these categories. Third, they investigate the used dataset in the proposed period. According the study, KDD99 has been used nearly 60% of the published studies. To expand the published review that included 55 articles, our study reviews 149 articles (Section 5), and includes more statistics.

[14] reviewed usage of swarm intelligence techniques in IDS. From these methods, ant colony optimization, ant colony clustering and particle swarm optimization has been compared in their review. Only descriptive statistic included in their study was performance comparison of swarm intelligence techniques in IDS.

[9] have reviewed intersection of feature selection and intelligent algorithms in Intrusion Detection. For feature selection, gradually feature removal method, modified mutual information-based feature selection algorithm, CRF-based feature selection, and wrapper based genetic feature selection methods have been compared. Regarding to classification techniques, [9] compares neural networks, genetic algorithms, fuzzy sets, rough sets, Neuro-Fuzzy, fuzzy-genetic algorithms and particle swarm optimization. They did not give any statistics that compare reviewed methods in their review.

[25] surveyed artificial immune systems in IDS. They reviewed concepts antibody/antigen encoding, generation algorithm, evolution algorithm but did not provide any statistics about the reviewed articles.

[8] surveyed evolutionary and swarm intelligence algorithms in network intrusion detection using DARPA and KDD99. They investigated usage of genetic algorithms, genetic programming, ant colony optimization and swarm optimization for different stages of IDS. In their study, the authors presented few descriptive statistics for evaluating the reviewed articles. First statistics is commonly used fitness functions, second statistics is articles' dataset usage. Third statistics is the applied algorithm, and the last statistics is detection rate of the applied algorithm.

Although the published and reviewed studies show IDS and Machine Learning is an active research topic in IDS, and KDD99 is the most used dataset, they do not provide enough statistics that shows how these methods are applied to KDD99. This study tries to present more comprehensive study to find satisfactory answers to mentioned questions by giving more statistics and checklist for guidance.

3 Datasets: DARPA, KDD99, and NSL-KDD

Figure 3 and Table 1 give overall summary for related datasets in this study —DARPA, KDD99, and NSL-KDD. DARPA is a base raw dataset. KDD99 is the feature extracted version of DARPA dataset. NSL-KDD is the duplicates removed and size reduced version of KDD99 dataset. Dataset statistics extracted from reviewed articles are given in Section 5.

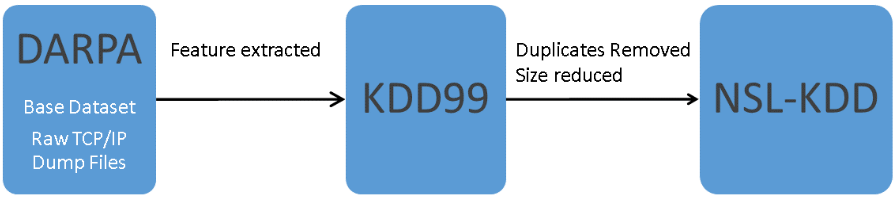

Fig. 3 The relation between main and extracted datasets. KDD99 is created from DARPA, NSL-KDD is created from KDD99.

Table 1 DARPA, KDD99, NSL-KDD Datasets Information

Name	Training Size	Testing Size	Note
DARPA 99	6.591.458 kb (6.2gb)	3.853.522 kb (3.67gb)	Base Dataset. Raw TCP/IP Dump files
KDD99	4898431	311029	Features extracted and preprocessed for machine learning
NSL-KDD	125973	22544	Duplicates removed, size reduced

3.1 DARPA Dataset

First DARPA-sponsored IDS-event was accomplished by MIT Lincoln LAB in 1998 [6]. In this DARPA event, an attack scenario to Air-Force base is simulated. One year later, in 1999, they repeated the same event [16] with improvements suggested by computer security community [17]. DARPA dataset consists of host and network dataset files. Host dataset, IDS bag, is small dataset that contains system calls, and is less used than its network counter part. Network dataset consists of seven weeks of raw TCP/IP dump files. Since DARPA dataset consists of raw files, researchers need to extract features from these files to use them in machine learning algorithms. First two weeks were attack free; therefore, it is suitable for training anomaly detection algorithms. Next five weeks, various attack was used against simulated air-force base, [13]. KDD99 dataset was created from DARPA network dataset files by Lee and Stolfo [15] in this competition.

3.2 KDD99 Dataset

Lee and Stolfo [15], one of the participating teams of the DARPA event, gave their feature extracted and preprocessed data to Knowledge Discovery and Data Mining (KDD) yearly competition [11]. Pfahringer [20] won KDD 99 competition using mixture of bagging and boosting. Most articles compare their results with winner's result [20]. KDD99 can be easily used in machine learning dataset; therefore, it is much more used in IDS and general research than DARPA dataset.

KDD99 has following characteristics:

1. It has two week's of attacks-free instances and five week's of attack instances, making it suitable for anomaly detection.
2. Output classes are divided to 5 main categories: These are DOS (Denial of Service), Probe, R2L (Root 2 Local), U2R (User 2 Root) and Normal.
3. Dataset contains 24 attack types in training and 14 more attack types in testing for total of 38 attacks. These 14 new attacks theoretically test IDS capability to generalize to unknown attacks. At the same time, it is hard for machine learning based IDS to detect these 14 new attacks [21].
4. It is heavily imbalanced dataset to attack instances. Approximately 80% percent of flow is attack traffic (3925650 attack instances in total 4898430 instances). Normally, typical network contains approximately 99.99% percent of normal instances. KDD99 violates this principle. Most articles needs to re-sample dataset to conform to typical network normality assumption, particularly anomaly detection articles.
5. U2R and R2L attacks are rare in KDD99 (Table 2).
6. Duplicate records in both training and testing datasets bias results for frequent DOS attacks and normal instances.
7. KDD99 is a large dataset for most machine learning algorithms; therefore, most studies use a small percentage of it.

Table 2 KDD99 Attack Distribution

	Training Size	(%)	Test Size	(%)
Normal	972781	19.85	60593	19.48
DOS	3883390	79.27	231455	74.41
Probe	41102	00.83	4166	01.33
U2R	52	00.001	245	00.07
R2L	1106	00.02	14570	04.68
Total	4898431	100	311029	100

KDD99's numerous shortcomings with respect to IDS is well documented in literature, [22, 19, 18, 4, 3].

3.3 NSL-KDD Dataset

To reduce deficiencies of KDD99 dataset for machine learning algorithms, Tavallaee et al. [23] introduced NSL-KDD dataset. NSL-KDD has been generated by removing redundant and duplicate instances, also by decreasing size of dataset. Since it is a re-sampled version of KDD99, IDS deficiencies remain in NSL-KDD.

4 General Machine Learning Work Flow Using KDD99

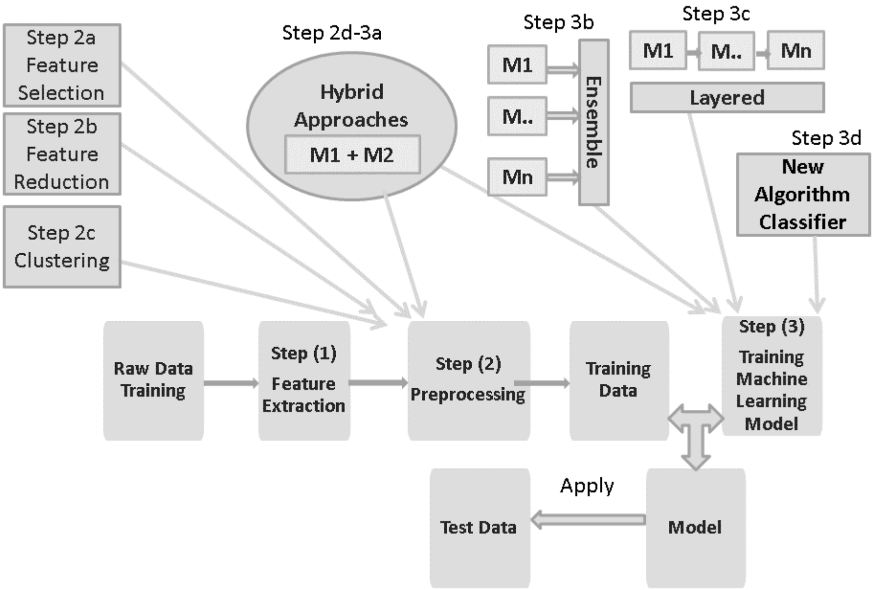

Fig. 4 General Machine Learning Flow Chart. Almost all of the reviewed articles make their contribution in steps 2a-2d and 3a-3d. Table 3 shows article counts for these contributions.

Figure 4 shows general machine learning work flow using any dataset. This work flow contains 3 main steps relevant to our discussion. These are step (1) feature extraction, step (2) preprocessing, and (step 3) training machine learning model. Normally, feature extraction step (1) is the most important step in machine learning. As KDD99 is a feature extracted dataset, this step is unnecessary.

Most reviewed studies made their contributions to preprocessing step (2) or training machine learning model step (3). For Step (2) preprocessing, reviewed articles used 4 different techniques: (2a) Feature Selection, (2b) Feature Reduction, (2c) Clustering, and (2d) Hybrid Approaches. Feature selection (2a) is using various algorithms to reduce number of existing 41 features. Feature

Table 3 Evaluating the reviewed articles regarding to machine learning model Figure 4

Contribution(Novelty)	Article Count	Figure 3
Hybrid	50	(2d-3a)
New Classifier Algorithm	45	(3d)
Feature Reduction	38	(2b)
Feature Selection	34	(2a)
New Anomaly Detection Algorithm	33	(3d)
New Optimization Algorithm	25	(2d-3a)
Layered	23	(3c)
New Clustering Algorithm	19	(2c)
Ensemble	14	(3b)
Agent Based	12	
Data Streams	7	

Transform (2b) is to change feature space of dataset to another space. For example, principal component analysis is a popular choice among reviewed studies(Table 6). Clustering (2c) is reduce features or instances using a clustering algorithm, for example k-means clustering. Hybrid Approaches (2d-3a) is using combination of two different algorithms for preprocessing or training machine learning model step. Most of the time, a feature selection/reduction/machine learning algorithm is hybridized with an optimization algorithm (for example: particle swarm optimization).

For Step (3) training machine learning model, reviewed articles used 4 different type of techniques: (3a) Hybrid Approaches, (3b) Ensemble, (3c) Layered, and (3d) New Algorithm Classifier. An example of Step (3a) Hybrid Approaches is training a neural networks with genetic algorithms instead of back propagation. Ensemble approach, Step (3b), is a parallel combination of different machine learning algorithms. Layering, Step (3c), is a serial combination of different machine learning algorithms. New Algorithm Classifier, Step (3d), means the applied algorithm may be entirely new or used the first time in IDS.

According to given work flow, contributions of the most reviewed articles may be more than two. For example, using a new optimization algorithm for feature selection and classification is counted as both feature selection and hybrid in this review. Also, using principal component analysis for feature reduction and using optimization algorithm to train a classifier is counted as both hybrid and feature reduction. Table 3 shows categorization of articles according to the work flow.

5 KDD99 Descriptive Statistics

Different from previous review studies, we present more descriptive statistics to evaluate published studies in focused period. Therefore, the following statistics has been extracted from the reviewed 149 studies:

1. Classification output classes

2. Training and Testing Dataset Usage
3. Cross Validation
4. Dataset sizes used in training and testing machine learning algorithms
5. Applied algorithms in proposed method
6. Classifiers used for comparison
7. Software Toolbox Usage
8. Other Datasets used in Reviewed Studies
9. Performance Metrics used in Experiments
10. IDS vs Not IDS
11. Main IDS Type according to study

These descriptive statistics are presented using figures and tables and has been discussed in detail.

5.1 Classification Output Classes in the Reviewed Studies

The output classes can be binary or multi classes when machine learning algorithms are applied to the KDD99. Table 4 shows output classes in reviewed articles. Multi class 5 are DOS,Probe,Normal,U2R and R2L as explained in Section 3.2. Multi Class X selects subset of 23 classes of KDD99, for example 7 attacks and normal and give results for 8 output. These studies are not comparable to other studies.

Table 4 Comparison of the published studied based on classification output classes.

Classification Output	Article Count
Binary (Attack/Normal)	124
Multiclass 5 (DOS/Probe/U2R/R2L/Normal)	49
No Binary: Gives other result	9
Multi Class X (Subset of 23)	22

5.2 Training and Testing Dataset Usage

Normally, in machine learning studies, datasets should be divided to training and testing datasets. Machine learning algorithms should be trained on training dataset and be tested on test dataset that is entirely separate from training datasets. Considering this usage, DARPA, KDD99 and NSL-KDD datasets contains two parts, training and testing. As mentioned before these two parts have different attacks and different probability distributions. Training a machine learning algorithm in a subset of KDD99 training dataset; then,

Table 5 Confusion Matrix for Training and Test Set Usage. Normally, only diagonal of matrix should have values, but most of the reviewed studies use KDD99 training dataset for both testing and training purposes.

		KDD99	
		Training	Test
Reviewed Study	Training	146	5
	Test	113	38

testing trained model in another subset gives optimistic results. Generally, machine learning algorithms should be trained on KDD99 training dataset and tested on KDD99 testing dataset.

Table 5 shows training and testing dataset usage in reviewed articles. Most reviewed articles (146) used KDD99 training dataset for training; but, 5 articles behaved differently. These 5 articles either merged training and testing dataset then re-sampled or used training dataset for testing purposes. The main reason for this application is to reduce difference between training and testing dataset. Only about 23% of reviewed articles (38) used real testing dataset. Others (113) used KDD99 training dataset as both testing and training purposes. This application provides optimistic results for these studies. Table 5 shows that most of the literature have used re-sampled version of KDD99 training dataset for both training and testing.

5.3 Cross Validation

K-fold cross validation is one of the suggested techniques in training machine learning models. Among the reviewed 149 studies, only 32 (%21) studies applied cross validation, while 117 (%79) studies did not apply cross validation.

5.4 Dataset sizes used in training and testing machine learning algorithms

In this review, 12 articles claimed KDD99 is a large dataset for machine learning research and used smaller subset of full dataset. Figure 5 and Figure 6 shows that training and testing dataset usage is skewed to small sizes. That is, most articles worked with small dataset sizes in reviewed studies. The smallest 10 training datasets contains 200 to 1000 instances, while smallest 10 testing datasets contains 80 and 1000 instances. These numbers are small compared to full size (4.9 Million instances) of KDD99 dataset. Using very small dataset sizes may be unacceptable from viewpoint of statistical analysis.

Figure 6 shows testing dataset sizes used in reviewed articles. Only usage of less than full size (311029 instances) of testing dataset are shown in Figure 6. About 15 study used larger numbers for testing, and all of them were data stream studies.

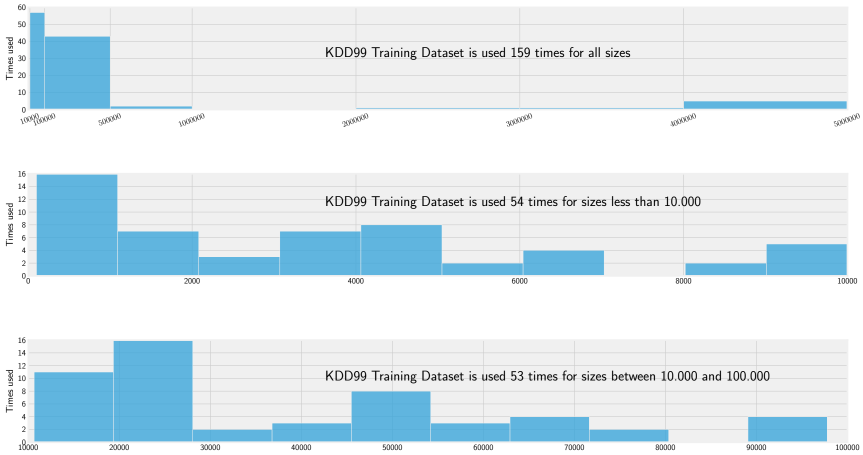

Fig. 5 KDD99 Training Usage Sizes. Most of the usage is with low sizes.

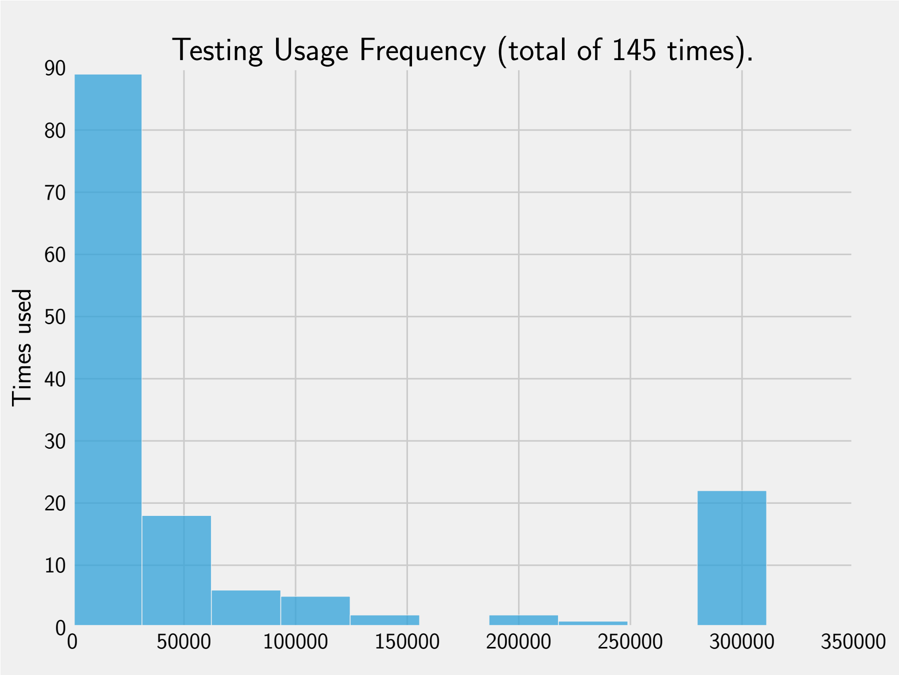

Fig. 6 KDD99 Testing Usage Sizes. Most of the low size usages comes from resampling of KDD99 Training dataset.

5.5 Applied algorithms in proposed method

Various algorithms have been used in KDD99 based IDS research. Table 6 shows algorithms that are used in proposed methods in reviewed studies. If a classifier, for example support vector machines, is used for comparison purposes, it is included in Table 7.

Table 6 Most used Algorithms in Proposed Methods.

Name	**Article Count**
Support Vector Machines	24
Decision Tree	19
Genetic Algorithm	16
Principal Component Analysis	13
Particle Swarm Optimization	9
k-nearest neighbors	9
k-means clustering	9
Naive Bayes	9
NeuralNetworks(MultiLayerPerceptron)	8
Genetic Programming	6
Rough Sets	6
Bayesian Network	5
Random Forest	5
Artificial Immune System	5
Fuzzy Rules Mining	4
NeuralNetworks(SelfOrganizingMap)	4

5.6 Classifiers used for comparison

Generally, most studies compare their results with other methods in their experiments. In KDD99 based machine learning research, most comparisons are made against other classifiers. Table 7 shows classifiers used for comparison in the reviewed studies. Some rare articles in this review were not machine learning articles, even though they use KDD99. These rare articles did not compare their suggested approaches with other classifiers. Also, some articles that proposed new methods for IDS, have not compared their proposed method with other classifiers. For these two types of articles, Table 7 includes *None*. For *Literature* entries, some articles did not compare their methods and their datasets using software toolboxes but only reported literature results. In our opinion, all machine learning IDS articles should use software toolboxes (Table 8) to compare their methods with common methods instead of only reporting literature results. Main reason for this advice is science reproducibility since every article is a bit different (sampling strategy, randomize seed, and different sizes for datasets).

5.7 Software Used in Reviewed Studies

Many software toolboxes has been used in IDS studies. Table 8 summarizes software used to implement or compare algorithms in the articles. Most arti-

Table 7 Classifiers used for comparison in Experiments.

Classifier	Article Count
Support Vector Machines	33
Naive Bayes	31
None (Not compared with other methods)	21
Decision Tree(J48)	21
k-nearest neighbors	20
Literature (results are given without experimental comparison)	19
Decision Tree	14
NeuralNetworks(MultiLayerPerceptron)	14
Bayesian Network	13
Random Forest	10
NeuralNetworks(SelfOrganizingMap)	7
NeuralNetworks(RadialBasisFunction)	7
K-Means	6
Rule Based Learner(JRipper)	5
Adaboost	5
Naive Bayes Tree	4
Decision Tree(C4.5)	4
PART	4
Decision Tree(CART)	4
Bagging	3
Random Tree	3

cles (78) did not give any information about applied software. This restricts reproducibility of applied method. Based on Table 8, Weka is widely used for classifier comparison even if it is not used for implementation. Matlab and Libsvm are also used for comparison. Most of the proposed methods are implemented using general purpose programming languages. As a remarkable note, although Python (2) and R (1) are touted as language of data science and machine learning [12], they were among the least used tools.

Table 8 Software used in Reviewed Articles. Weka, Matlab, and Libsvm are used for comparison purposes. General purpose programming languages are used for implementation. Software that are used less than two is not included.

Software Tool/Package	Article Count
No Information(software used is unclear)	78
Weka	34
Matlab	26
Libsvm	9
Java	7
C++	5
CSharp	3
Pascal;Hadoop;Python;MOA	2 for each

5.8 Different datasets used in reviewed studies

In addition to KDD99, different datasets were also used in the reviewed articles, Table 9. Non IDS datasets in this review show that KDD99 is used as just another dataset in some studies. NSL-KDD is re-sampled version of KDD99 as it is explained in Section 3.3. Some studies used both NSL-KDD and KDD99, while others used only NSL-KDD dataset. Other IDS datasets —ISCX, Kyoto — are only used 3 times, about 2% of all articles in Table 9. This shows lack of recent IDS dataset.

Table 9 Most used Datasets. * denotes IDS datasets. Datasets that are used less than three is not included.

Dataset Name	Article Count
KDD99*	133
NSL-KDD*	23
DARPA*	9
Iris	8
Glass	5
Breast Cancer	5
Synthetic Data	5
Poker Hand	5
Image Segmentation	3
ISCX*	3
Wine	3
Kyoto*	3

5.9 Performance Metrics Used in Reviewed Studies

Various performance metrics can be used to evaluate to machine learning algorithms. Table 10 summarizes which metrics are provided in the 149 articles reviewed. Detection rate is most consistent metric provided; although some articles fail to provide this metric. For example, some articles gave figures for their detection rate but did not give a number; therefore, reader has to guess about its value. Other articles gave 5-class detection rates but did not give overall detection rate for comparison. If researcher would like to compare given result with other articles, it is often impractical since dataset sizes differ greatly from article to article. Some articles gave detection rate by class but failed to provide number of class instances therefore it is impossible to get single detection rate for attack versus normal.

Some articles did not present information about used testing dataset. Machine learning algorithms get different results in KDD99 train and testing dataset as mentioned in Section 3.2. Therefore; it is important for articles that use KDD99 to indicate if they used training or testing dataset of KDD99.

Other performance metrics differ widely in our reviewed articles. Computational Complexity metrics were not given in most of the articles. Also training

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1954v1 | CC-BY 4.0 Open Access | rec: 14 Apr 2016, publ: 14 Apr 2016

Table 10 Performance Metrics Used. Usage of performance metrics are highly irregular. Some articles does not give any metric(*).

Performance Metric	Article Count
Detection Rate	134
False Positive(False Alarm)	70
Training Time	44
Testing Time	28
ROC-Curve	24
False Negative	22
Confusion Matrix (5 class)	20
True Positive	20
Error Rate	13
Precision	13
F-Measure	13
Recall	12
True Negative	11
Number of Selected Features	10
Correlation Coefficient	9
Cost Per Example	9
ROC-Area Under Curve	8
Sensitivity	7
Specificity	7
Root Mean Square Error	6
None*	5
Memory Usage	5

time was given 44 (%29) times , and testing time was given 28 (% 18) times. Considering importance of these two metrics, their usage is low.

Generally, we suggest that following metrics are provided in the results. If the study is multi-class, multi-class versions and binary versions should be provided together. (1) Detection Rate, (2) Confusion matrix, (3) Training Time, (4) Testing Time, (5) Computational Complexity for newly proposed methods.

5.10 Main IDS Type according to study

Figure 7 shows which IDS methodologies are used in the collected articles. Total count is more than 149, since most articles use more than one methodology. KDD99 is a popular choice for both machine learning and anomaly detection studies.

5.11 IDS vs Not IDS Studies

Figure 8 shows how many articles claim that they are IDS studies among the reviewed articles. Even though, IDS articles are majority, number of Non IDS articles shows that KDD99 is also widely used in other domains.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1954v1 | CC-BY 4.0 Open Access | rec: 14 Apr 2016, publ: 14 Apr 2016

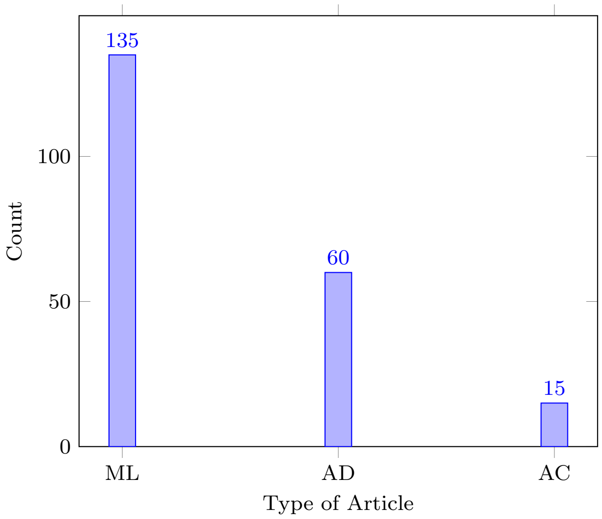

Fig. 7 Article Count by Methodology (ML:Machine Learning, AD:Anomaly Detection, AC:Alert Correlation)

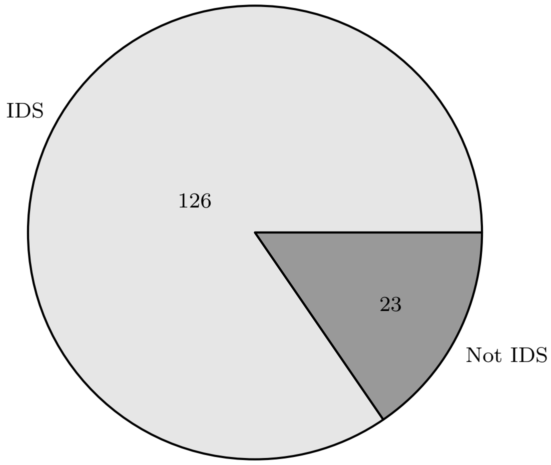

Fig. 8 IDS and Not IDS articles in 2010–2015. KDD99 is mostly used in IDS studies; but, some studies use it also, especially feature selection and data streams studies.

6 A Suggested Checklist for Avoiding most Common Mistakes in IDS and MLR

After evaluating 149 studies from the SCI-indexed 65 journals, the authors suggest a checklist for researchers who want to apply a machine learning or IDS method using KDD99 or other large dataset. The suggested checklist has been provided considering common mistakes and strengths points of the reviewed studies. The suggested instructions could be useful for future studies in a similar area. The instructions are as follows:

- Point out training and test dataset clearly.
- If the target dataset is KDD99, identify if full dataset or a portion of dataset is used.
- Specify train, test dataset and validation dataset sizes in a table.
- Increase reproducibility of the study by giving software package, re-sampling strategy, and random seeds.
- To evaluate the classification result, provide confusion matrix, detection rate, training time and testing time.
- Compare the result of the proposed approach with other most used methods.
- Determine the number of output classes. For KDD99 using 5 or 23 classes will be preferred.

7 Conclusion

In the proposed study, 149 recent studies using KDD99 dataset between 2010 and 2015 have been reviewed. A different review process is followed from previous reviews in the same area. Instead of finding the major contributions to the area, descriptive statistics are extracted. Review results show the following findings. First, even though KDD99 is an 17-years-old dataset, it is still widely used in IDS and machine learning research. Second, decision tree derivatives and support vector machines are most applied algorithms. Third, Weka and Matlab are the most used software toolbox, even though most studies did not give information about software usage. Fourth, detection rate is the most used performance metric to show classification quality. Additionally, considering common errors and strengths of the reviewed works, a checklist has been suggested to improve the research quality in similar areas.

References

References

1. Allen WH (2007) Mixing wheat with the chaff: Creating useful test data for ids evaluation. IEEE Security and Privacy 5:65–67, DOI 10.1109/MSP.2007.92, URL http://portal.acm.org/citation.cfm?id=1308457.1309257
2. Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A, Benitez J, Herrera F (2014) A review of microarray datasets and applied feature selection methods. Information Sciences 282:111 – 135, DOI http://dx.doi.org/10.1016/j.ins.2014.05.042, URL http://www.sciencedirect.com/science/article/pii/S0020025514006021

Table 11 Journals and Article Counts

Journal Name	Article Count	Journal Name	Article Count
Expert System With Applications	22	Future Generation Computer Systems	1
Security and Communication Networks	12	Artificial Intelligence Review	1
Applied Soft Computing	7	Computing and Informatics	1
Knowledge-Based Systems	7	Soft Computing	1
The Scientific World Journal	6	IEEE Systems Journal	1
Information Sciences	5	International Journal of Systems Science	1
Applied Intelligence	5	The Computer Journal	1
Neurocomputing	5	Journal Of Intelligent Information Systems	1
Neural Computing and Applications	4	Wireless Personal Communications	1
Computer Communications	4	Discrete Dynamics in Nature and Society	1
Engineering Applications Of Artificial Intelligence	3	Iranian Journal of Fuzzy Systems	1
Pattern Recognition	3	IETE Technical Review	1
International Journal Of Computational Intelligence Systems	3	Simulation Modelling Practice and Theory	1
International Journal of Computer Science and Network Security	3	Telecommunication Systems	1
Intelligent Automation & Soft Computing	2	China Communications	1
IEEJ Transactions On Electrical and Electronic Engineering	2	Turkish Journal Of Electrical Engineering And Computer Sciences	1
IEEE Transactions on Knowledge and Data Engineering	2	Neural Processing Letters	1
Mathematical Problems in Engineering	2	IEEE Transactions on Computers	1
International Journal of Innovative Computing Information and Control	2	IEEE Transactions on Smart Grid	1
Genetic Programming and Evolvable Machines	2	Applied Artificial Intelligence	1
IETE Journal of Research	2	Defence Science Journal	1
The International Arab Journal of Information Technology	2	The Journal of Supercomputing	1
Journal of Network and Computer Applications	2	International Journal of Fuzzy Logic and Intelligent Systems	1
Applied Mathematics and Information Sciences	1	Data Mining and Knowledge Discovery	1
PLOS One	1	Journal of The Faculty of Engineering and Architecture of Gazi University	1
IEEE Transactions on Cybernetics	1	Knowledge and Information Systems	1
IEEE Transactions on Dependable and Secure Computing	1	Journal of Parallel and Distributed Computing	1
IEEE Transactions on Systems Man And Cybernetics	1	Computational Intelligence and Neuroscience	1
Journal of Information Science and Engineering	1	Iranian Journal of Science and Technology Transactions of Electrical Engineering	1
IEEE Transactions on Neural Networks and Learning Systems	1	Acta Polytechnica Hungarica	1
IEEE Transactions On Parallel And Distributed Systems	1	Cybernetics and Information Technologies	1
Journal of Network and Systems Management	1	Artificial Intelligence	1
Journal of Advanced Research	1		

3. Brugger S (2007) Kdd cup 99 dataset (network intrusion) considered harmful
4. Brugger ST, Chow J (2005) An assessment of the darpa ids evaluation dataset using snort
5. Catania CA, Garino CG (2012) Automatic network intrusion detection: Current techniques and open issues. Computers & Electrical Engineering 38(5):1062 – 1072, DOI http://dx.doi.org/10.1016/j.compeleceng.2012.05.013, URL http://www.sciencedirect.com/science/article/pii/S0045790612001073, special issue on Recent Advances in Security and Privacy in Distributed Communications and Image processing
6. Cunningham RK, Lippmann RP, Fried DJ, Garfinkel SL, Graf I, Kendall KR, Webster SE, Wyschogrod D, Zissman MA (1999) Evaluating intrusion detection systems without attacking your friends: The 1998 darpa intrusion detection evaluation. Tech. rep., MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB
7. Denning DE (1987) An intrusion-detection model. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING 13(2):222–232
8. Elsayed S, Sarker R, Essam D (2015) Survey of uses of evolutionary computation algorithms and swarm intelligence for network intrusion detection. International Journal of Computational Intelligence and Applications 14(04):1550,025, DOI 10.1142/S146902681550025X, URL http://www.worldscientific.com/doi/abs/10.1142/S146902681550025X, http://www.worldscientific.com/doi/pdf/10.1142/S146902681550025X
9. Ganapathy S, Kulothungan K, Muthurajkumar S, Vijayalakshmi M, Yogesh P, Kannan A (2013) Intelligent feature selection and classification techniques for intrusion detection in networks: a survey. EURASIP Journal on Wireless Communications and Networking 2013(1):271, DOI 10.1186/1687-1499-2013-271, URL http://dx.doi.org/10.1186/1687-1499-2013-271

10. Hubballi N, Suryanarayanan V (2014) False alarm minimization techniques in signature-based intrusion detection systems: A survey. Computer Communications 49:1 – 17, DOI http://dx.doi.org/10.1016/j.comcom.2014.04.012, URL http://www.sciencedirect.com/science/article/pii/S0140366414001480
11. KDD (1999) Intrusion detector learning
12. kdnuggets (2015) New poll: Primary programming language for analytics, data mining, data science. http://goo.gl/ESGonu
13. Kendall K (1999) A database of computer attacks for the evaluation of intrusion detection systems. Master's thesis, MIT - Massachusetts Institute of Technology
14. Kolias C, Kambourakis G, Maragoudakis M (2011) Swarm intelligence in intrusion detection: A survey. Computers and Security 30(8):625–642, DOI 10.1016/j.cose.2011.08.009, URL http://dx.doi.org/10.1016/j.cose.2011.08.009
15. Lee W, Stolfo SJ (2000) A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security 3:227–261, DOI http://doi.acm.org/10.1145/382912.382914, URL http://doi.acm.org/10.1145/382912.382914
16. Lippmann R, Fried D, Graf I, Haines J, Kendall K, McClung D, Weber D, Webster S, Wyschogrod D, Cunningham R, Zissman M (2000) Evaluating intrusion detection systems: the 1998 darpa off-line intrusion detection evaluation. In: DARPA Information Survivability Conference and Exposition, 2000. DISCEX '00. Proceedings, vol 2, pp 12 –26 vol.2, DOI 10.1109/DISCEX.2000.821506
17. Lippmann R, Haines JW, Fried DJ, Korba J, Das K (2000) The 1999 darpa off-line intrusion detection evaluation. Computer Networks 34:579–595, DOI 10.1016/S1389-1286(00)00139-0, URL http://portal.acm.org/citation.cfm?id=361116.361124
18. Mahoney M, Chan P (2003) An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection. In: Recent Advances in Intrusion Detection, Springer, pp 220–237
19. McHugh J (2000) Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Transactions on Information and System Security 3(4):262–294
20. Pfahringer B (2000) Winning the kdd99 classification cup: bagged boosting. SIGKDD Explor Newsl 1:65–66, DOI http://doi.acm.org/10.1145/846183.846200, URL http://doi.acm.org/10.1145/846183.846200
21. Sabhnani M, Serpen G (2004) Why machine learning algorithms fail in misuse detection on kdd intrusion detection data set. Intell Data Anal 8:403–415, URL http://portal.acm.org/citation.cfm?id=1293805.1293811
22. Sommer R, Paxson V (2010) Outside the closed world: On using machine learning for network intrusion detection. In: Proceedings of the 2010 IEEE Symposium on Security and Privacy, IEEE Computer Society, Washing-

ton, DC, USA, SP '10, pp 305–316

23. Tavallaee M, Bagheri E, Lu W, Ghorbani AA (2009) A detailed analysis of the kdd cup 99 data set. In: Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, IEEE Press, Piscataway, NJ, USA, CISDA'09, pp 53–58, URL http://portal.acm.org/citation.cfm?id=1736481.1736489
24. Tsai CF, Hsu YF, Lin CY, Lin WY (2009) Intrusion detection by machine learning: A review. Expert Systems with Applications 36(10):11,994 – 12,000, DOI DOI:10.1016/j.eswa.2009.05.029, URL http://www.sciencedirect.com/science/article/pii/S0957417409004801
25. Yang H, Li T, Hu X, Wang F, Zou Y (2014) A survey of artificial immune system based intrusion detection. The Scientific World Journal 2014:11, DOI 10.1155/2014/156790, URL http://dx.doi.org/10.1155/2014/156790