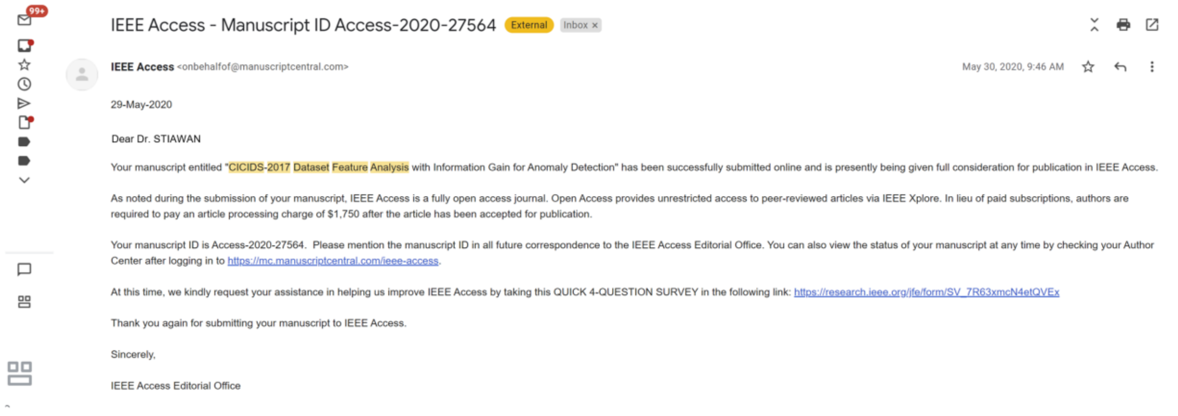


IEEE Access - Manuscript ID Access-2020-27564

IEEE Access <onbehalfof@manuscriptcentral.com>

May 30, 2020, 9:46 AM

29-May-2020

Dear Dr. STIAWAN

Your manuscript entitled "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection" has been successfully submitted online and is presently being given full consideration for publication in IEEE Access.

As noted during the submission of your manuscript, IEEE Access is a fully open access journal. Open Access provides unrestricted access to peer-reviewed articles via IEEE Xplore. In lieu of paid subscriptions, authors are required to pay an article processing charge of $1,750 after the article has been accepted for publication.

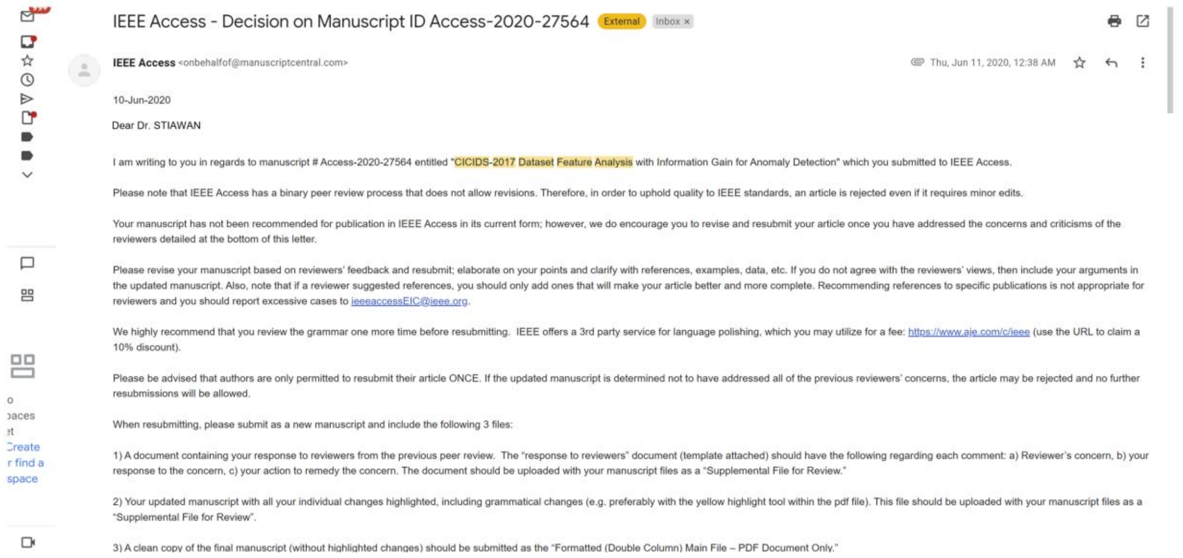
Your manuscript ID is Access-2020-27564. Please mention the manuscript ID in all future correspondence to the IEEE Access Editorial Office. You can also view the status of your manuscript at any time by checking your Author Center after logging in to https://mc.manuscriptcentral.com/ieee-access.

At this time, we kindly request your assistance in helping us improve IEEE Access by taking this QUICK 4-QUESTION SURVEY in the following link: https://research.ieee.org/jfe/form/SV_7R63xmcN4etQVEx

Thank you again for submitting your manuscript to IEEE Access.

Sincerely,

IEEE Access Editorial Office



IEEE Access - Decision on Manuscript ID Access-2020-27564

IEEE Access <onbehalfof@manuscriptcentral.com>

Thu, Jun 11, 2020, 12:38 AM

10-Jun-2020

Dear Dr. STIAWAN

I am writing to you in regards to manuscript # Access-2020-27564 entitled "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection" which you submitted to IEEE Access.

Please note that IEEE Access has a binary peer review process that does not allow revisions. Therefore, in order to uphold quality to IEEE standards, an article is rejected even if it requires minor edits.

Your manuscript has not been recommended for publication in IEEE Access in its current form; however, we do encourage you to revise and resubmit your article once you have addressed the concerns and criticisms of the reviewers detailed at the bottom of this letter.

Please revise your manuscript based on reviewers' feedback and resubmit; elaborate on your points and clarify with references, examples, data, etc. If you do not agree with the reviewers' views, then include your arguments in the updated manuscript. Also, note that if a reviewer suggested references, you should only add ones that will make your article better and more complete. Recommending references to specific publications is not appropriate for reviewers and you should report excessive cases to ieeeaccessEIC@ieee.org.

We highly recommend that you review the grammar one more time before resubmitting. IEEE offers a 3rd party service for language polishing, which you may utilize for a fee: https://www.aje.com/c/ieee (use the URL to claim a 10% discount).

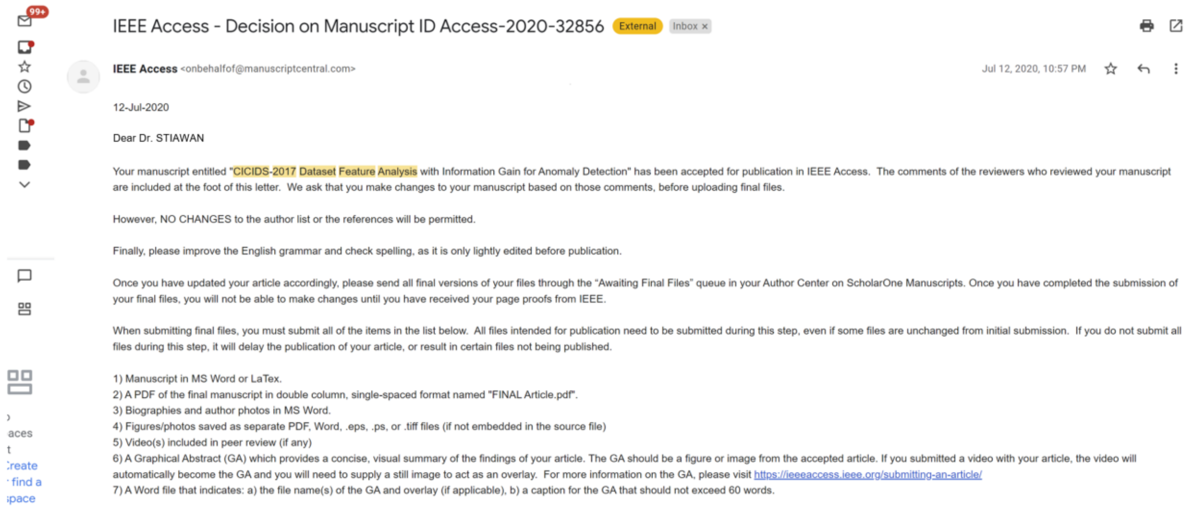
Please be advised that authors are only permitted to resubmit their article ONCE. If the updated manuscript is determined not to have addressed all of the previous reviewers' concerns, the article may be rejected and no further resubmissions will be allowed.

When resubmitting, please submit as a new manuscript and include the following 3 files:

1) A document containing your response to reviewers from the previous peer review. The "response to reviewers" document (template attached) should have the following regarding each comment: a) Reviewer's concern, b) your response to the concern, c) your action to remedy the concern. The document should be uploaded with your manuscript files as a "Supplemental File for Review."

2) Your updated manuscript with all your individual changes highlighted, including grammatical changes (e.g. preferably with the yellow highlight tool within the pdf file). This file should be uploaded with your manuscript files as a "Supplemental File for Review".

3) A clean copy of the final manuscript (without highlighted changes) should be submitted as the "Formatted (Double Column) Main File – PDF Document Only."



IEEE Access - Decision on Manuscript ID Access-2020-32856

IEEE Access <onbehalfof@manuscriptcentral.com>

Jul 12, 2020, 10:57 PM

12-Jul-2020

Dear Dr. STIAWAN

Your manuscript entitled "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection" has been accepted for publication in IEEE Access. The comments of the reviewers who reviewed your manuscript are included at the foot of this letter. We ask that you make changes to your manuscript based on those comments, before uploading final files.

However, NO CHANGES to the author list or the references will be permitted.

Finally, please improve the English grammar and check spelling, as it is only lightly edited before publication.

Once you have updated your article accordingly, please send all final versions of your files through the "Awaiting Final Files" queue in your Author Center on ScholarOne Manuscripts. Once you have completed the submission of your final files, you will not be able to make changes until you have received your page proofs from IEEE.

When submitting final files, you must submit all of the items in the list below. All files intended for publication need to be submitted during this step, even if some files are unchanged from initial submission. If you do not submit all files during this step, it will delay the publication of your article, or result in certain files not being published.

1) Manuscript in MS Word or LaTex.
2) A PDF of the final manuscript in double column, single-spaced format named "FINAL Article.pdf".
3) Biographies and author photos in MS Word.
4) Figures/photos saved as separate PDF, Word, .eps, .ps, or .tiff files (if not embedded in the source file)
5) Video(s) included in peer review (if any)
6) A Graphical Abstract (GA) which provides a concise, visual summary of the findings of your article. The GA should be a figure or image from the accepted article. If you submitted a video with your article, the video will automatically become the GA and you will need to supply a still image to act as an overlay. For more information on the GA, please visit https://ieeeaccess.ieee.org/submitting-an-article/
7) A Word file that indicates: a) the file name(s) of the GA and overlay (if applicable), b) a caption for the GA that should not exceed 60 words.

Original Manuscript ID: Access-2020-27564

Original Article Title: "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection"

To: IEEE Access Editor

Re: Response to reviewers

Dear Editor,

Thank you for allowing a resubmission of our manuscript, with an opportunity to address the reviewers' comments.

We are uploading (a) our point-by-point response to the comments (below) (response to reviewers), (b) an updated manuscript with yellow highlighting indicating changes, and (c) a clean updated manuscript without highlights (PDF main document).

Best regards,

Deris Stiawan

Reviewer#1, Concern # 1: There is no discussion on the cost effectiveness of the FS or algorithms used. What is the computational complexity? What is the runtime? Please include such discussions. You can also use the big oh notation to show the computation complexity

Author response: We provide short analysis on the complexity of FS technique used in this paper.

Author action: We updated the manuscript by adding information on the computational complexity of the FS technique and the algorithms under sub-section III.C as follows.

This work chooses Information Gain as feature selection since it is a filtered-based technique which provides more stable sets of selected features due to its robust nature against overfitting. Overall, computational complexity of filter-based technique is $O(m \cdot n^2)$, where m is the number of training data, and n is number the of attributes/features. It is less as compared to embedded and wrapper-based techniques [55]. The complex nature of wrapper-based techniques creates the high risk of overfitting. Thus, using feature selection technique that produces significant, relevant, less number of features and less computational complexity will reduce the execution time of classification algorithms used in the anomaly/attack detection process.

Reviewer#1, Concern # 2: To have an unbiased view in the paper, there should be some discussions on the limitations the FS methods used

Author response: The authors thank for this valuable comment, it improves the paper.

Author action: We updated the manuscript by adding the discussion about Information Gain limitation in the last paragraph of Section II as follows.

On the other hand, despite many researchers using Information Gain as a feature selection technique, there are very limited discussions on how to determine the minimum weight or rank score from the Information Gain result. This score determines how much the features are relevant to the class label. Researchers in [18] and in [21] use a score feature above 0.4 and a score above 0.001, respectively. Meanwhile, research work in [28] considers the minimum weight score of 0.8. In contrast, researchers in [29] remove features one by one and apply the classifier algorithm to find the best accuracy. Such work is very time-consuming especially with a large number of features in the dataset.

Reviewer#1, Concern # 3: Analysis of the results is missing in the paper. There is a big gap between the results and conclusion. There should be the result analysis between these two sections. After comparing the methods, you have to be able to analyse the results and relate them to the structure of all algorithms. It would be interesting to have your thoughts on why the method works that way? Such analyses would be the core of your work where you prove your understanding of the reason behind the results. You can also link the findings to the hypotheses of the paper. Long story short, this paper requires a very deep analysis from different perspectives

Author response: Very appreciate your suggestion to make the paper solid.

Author action: We updated the manuscript with more analysis on the results in sub-section IV.D, page 8, as follows.

Implementation of the proposed Information Gain feature selection in the experiments yields ranked features according to their weight scores. Features with higher weight scores represent more relevant and significant features of an attack. Table 13 depicts the top four features with their scores resulted from the experiment. Thus, features with IDs 41, 13, 65, and 8 are the most relevant and significant features for detecting any attacks and appear in any of features subsets.

TABLE XIII

Top four Features resulted from the proposed Information Gain Implementation.

No	*Feat. ID*	*Feature Names*	*Weight*
1	*41*	*Packet Length Std*	*0,638*
2	*13*	*Total Length of Bwd Packets*	*0,612*
3	*65*	*Subflow Bwd Bytes*	*0,612*
4	*8*	*Destination Port*	*0,609*

Overall, RF, BN, RT and J48 classifiers are able to detect well the normal traffic, DoS/DDoS, Port Scan, Brute Force and Web attacks traffic using the features subsets of 35, 52, and 77. Literatures study supports this finding as the classifiers use robust decision tree learning algorithm.

For the case of Infiltration attack traffic detection, NB is able to detect with TPR value of 0.800 using features subsets of 22 and 35, and perfectly detect (with TPR value of 1.000) using features subsets of 52, 57 dan 77. The reason is, because significant features representing infiltration attack traffic appears in the features subsets of 52, 55, 77. Unfortunately, other classifiers; RF, BN, RT and J48 are unable to detect well the Infiltration attack traffic. The small amount of this type of attack traffic in the dataset may cause the bad performance of its detection. As mentioned in Subsection 4.A, CCIDS-2017 contains imbalanced data, which is a big challenge in detecting anomalies/attacks.

Similar to the case of Infiltration attack, all classifiers are not able to detect well the Web Attack traffic using features subset of 4. Then, only BN and NB classifiers are able to detect the Web Attack traffic using features subset of 15 with the TPR value of 0.993 and 0.829, respectively.

As for Bot Attack traffic detection, RF, BN, RT, and J48 are able to detect the traffic using certain features subsets, but with lower TPR values.

Furthermore, considering the Precision and Recall values, in general the five classifiers detect the traffic relatively well. Nevertheless, in some cases the classifiers produce NaN values. Those cases may happen because of the implementation of 10-Fold Cross Validation in the experiment, which divides the dataset into ten folds (data portion). As the amount of attack traffics for Infiltration, Bot and Web attacks are relatively small, thus, some folds do not contain those traffics. Therefore, it affects the ability to detect the attack during the training stage. Specifically, for the Infiltration attack traffic which has very small amount in the dataset.

Reviewer#1, Concern # 4: How do you ensure that the comparison between the methods is fair?

Author response: The experiments use the same dataset and same number of records for each feature group. Besides, we use the same data sampling technique (10-fold cross validation), which is commonly used by other researchers. Thus, we ensure that the comparison between the classifier algorithms/methods is fair enough.

Author action: No action on the manuscript.

Reviewer#1, Concern # 5: The FS method might be sensitive to the values of its main controlling parameter. How did you tune the parameters?

Author response: The Information Gain Feature Selection (IGFS) ranks the features based on their weight values and the minimum weight is determined manually using try and error approach. In this work, we propose to group the features according to the minimum weight values. Thus, we obtain groups of features, where each feature group will be having different number of features as shown in Table 5. Further, all feature groups will be validated by using five classifiers, so we can determine which feature group are effective enough to be used for attacks' types classification.

Author action: We updated the manuscript in sub-section III. C, last paragraph, page 4 as follows.

The Information Gain ranks the features based on their weight values and the minimum weight is determined manually using try and error approach. In this work, the researchers propose to group the features according to the minimum weight values. Thus, groups of features are obtained and each feature group will be having different number of features as shown in Table 5. Further, all feature groups will be validated by using five classifier algorithms, so we can determine which feature group are effective enough to be used for attacks' types classification.

Reviewer#2, Concern # 1: What is the main motivation of the work? How this work is different from the literature?

Author response: Previous works that use the same dataset and also use Information Gain FS do not mention the basis on how to determine the score value used for feature selection. Each researchers use different score value. In this paper, we investigate and analyze the ability of the Information Gain technique in determining relevant features for network traffic classification, especially for traffic with bigger number of features. We distribute the features into groups based on their minimum score values. Then each feature group is used as a filter for the five classifier algorithms; Random Forest, Bayes Network, Random Tree, Naive Bayes and J48 to perform anomaly/attack detection on the dataset. Then, the detection results are compared with the aim is to validate the significance and relevant of the selected feature groups. The more accurate the detection results the more significance and relevance the features group. Thus, we analyze the effect of weighted features resulted from the Information Gain technique against the anomaly/attack detection performance as well as to find the most significant and relevant features to be used to increase the performance of anomaly/attack detection.

Author action: We updated the manuscript by adding the following paragraph into Introduction Section, in para 5, page 2.

Previous works that use the same dataset and also use Information Gain feature selection technique do not mention the basis on how to determine the score value used for feature selection. Each researchers use different score value. In this paper, the authors investigate and analyze the ability of the Information Gain technique in determining relevant features for network traffic classification, especially for traffic with bigger number of features. The authors distribute the features into groups based on their minimum score values. Then each feature group is used as a filter for the five classifier algorithms; Random Forest, Bayes Network, Random Tree, Naive Bayes and J48 to perform anomaly/attack detection on the dataset. Then, the detection results are compared with the aim is to validate the significance and relevance of the selected feature groups. The more accurate the detection results the more significance and relevant the features group. Thus, the authors analyze the effect of weighted features resulted from the Information Gain technique against the anomaly/attack detection performance as well as to find the most significant and relevant features to be used to increase the performance of anomaly/attack detection.

Reviewer#2, Concern # 2: What are the complex features that are considered in the work? Are the features not available with the NSL KDD sets?

Author response:

In this work, we consider complex features that represent sophisticated attacks on modern network based on its traffic attributes. For examples, features that exist in CICIDS-2017 but are not available in NSL-KDD include: *Subflow Fwd Bytes* and *Total Length Fwd Package* which are required to detect Infiltration and Bot attack types. The *Bwd Packet Lenght Std* feature is required to detect the types of DDoS, DoS Hulk, DoE GoldenEye, and Heartbleed attacks. The *Init Win Fwd Bytes* feature is required to detect the types of Web-Attack, SSH-Patator, and FTP-Patator attacks. Whereas the *Min Bwd Package Length* feature and *Fwd Average Package Length* feature are required to recognize normal traffic.

Author action: We add discussion in second paragraph of sub-section III.A in page 3, as follows.

In this work, the authors consider complex features that represent sophisticated attacks on modern network based on its traffic attributes. For examples, features that exist in CICIDS-2017 but are not available in NSL-KDD include: Subflow Fwd Bytes and Total Length Fwd Package which are required to detect Infiltration and Bot attack types. The Bwd Packet Lenght Std feature is required to detect the types of DDoS, DoS Hulk, DoE GoldenEye, and Heartbleed attacks. The Init Win Fwd Bytes feature is required to detect the types of Web-Attack, SSH-Patator, and FTP-Patator attacks. Whereas the Min Bwd Package Length feature and Fwd Average Package Length feature are required to recognize normal traffic [58].

CICIDS-2017 has more complex types of attacks as presented in Table 2. The rational of choosing CICIDS-2017 dataset is to have a dataset that represents closely the current real world network traffic in our experiments.

Reviewer#2, Concern # 3: Experimental design diagram to be clearly visible.

Author response: Noted, It happened due to the copy paste from the original draft to the template.

Author action: We revised and re-draw the figure, please refer Figure 1 in page 4.

Reviewer#2, Concern # 4: Information gain is already used in the experiments in the literature. Authors to clearly state the contribution.

Author response: In this work, we conduct experiments to investigate the wellness of feature groups resulted from Information Gain feature selection technique and validate their significance and relevance to the anomaly/attack detection accuracy using five classifier algorithms.

Author action: We updated the manuscript in sub-section III. C, last paragraph, page 4 as follows.

The Information Gain ranks the features based on their weight values and the minimum weight is determined manually using try and error approach. In this work, the researchers propose to group the features according to the minimum weight values. Thus, groups of features are obtained and each feature group will be having different number of features as shown in Table 5. Further, all feature groups will be validated by using the five classifiers, so we can determine which feature groups are effective enough to be used for attacks' types classification.

Reviewer#2, Concern # 5: Relabeling is not indicated in the architecture / setup.

Author response: Thank you to pin point the missing important process in the experimental design. We fix the missing relabeling process in the design.

Author action: We edit the experimental design in Figure 1 by adding relabeling process. It is in the first box after the input, named "Remove Redundant Features & Relabeling" process. (Please refer to Figure 1 in page 4). And edit the first step in Experimental set up, in sub-section III.C to include the relabeling process.

Reviewer#2, Concern # 6: How this setup would work in the standard data such as NSL KDD?

Author response: In principle, the set up can be implemented on the NSL-KDD dataset. As the set up applies ranking on the feature scores, and the NSL-KDD dataset consists of less number of features, thus, definitely, the set up will work well and produce the best grouping of features.

Author action: No action for this concern

Reviewer#2, Concern # 7: How the performance results provide the significance in the work?

Author response: We add analysis on this matter in Section IV.D in page 8 and updated the manuscript

Author action: We add analysis on this matter in Section IV.D in page 8 and updated the manuscript as follows.

Implementation of the proposed Information Gain feature selection in the experiments yields ranked features according to their weight scores. Features with higher weight scores represent more relevant and significant features of an attack. Table 13 depicts the top four features with their scores resulted from the experiment. Thus, features with IDs 41, 13, 65, and 8 are the most relevant and significant features for detecting any attacks and appear in any of features subsets.

Reviewer#2, Concern # 8: Authors may look at the following articles:

1. A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems, 2018.
2. Identifying and Benchmarking Key Features for Cyber Intrusion Detection: An Ensemble Approach, 2019.
3. Intrusion detection model using fusion of chi-square feature selection and multi class SVM, 2017.
4. Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers, 2019.
5. Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset, 2019.

Author response: The authors thank to the reviewer for pointing out missing important related works. As we focus on filter-based feature selection, we limit the literatures that discuss about the filter-based FS articles. The first suggested reference has already been in the reference list in the first draft of the manuscript (Reference list number 44).

Author action: We added the fifth suggested reference (reference list number [56]) and updated the Relevant Researches Section in page 3, in the last entry of Table 1 and before the last paragraph as follows.

Work in [56] combines the Synthetic Minority Oversampling Technique (SMOTE), Principal Component Analysis (PCA), and Ensemble Feature Selection (EFS) to improve the performance of AdaBoost-based IDS on the CICIDS 2017 Dataset. The authors claim that the combined method outperforms the SVM-based method with regards to accuracy, precision, recall and F1 Score.

Reviewer#3, Concern # 1: Page numbering is problematic. It always looks 9 after somewhere.

Author response: apologize for not being thorough in preparing the manuscript, need time to familiarize with the template

Author action: We fix the page numbering

Reviewer#3, Concern # 2: They didn't explain why they did some things. Why was 10-fold made? Why was the data set divided by 70% and 30%?

Author response: We provide explanations on the 10-folds and why we use 70%-30% data portion for training and testing data.

Author action: We updated the manuscript by adding an explanation of why 10-fold cross is used in sub-section IV.C in page 6, as follows.

The 10-fold cross-validation is used because it reduces computing time while maintains the performance of the classification algorithms in term of accuracy. Hence, the input dataset will be randomly divided into 10 folds with exactly the same size. For each of the 10 fold data, cross-validation will use 9 fold for training and 1 fold for testing. This process is repeated for 10 times until each fold becomes a test fold. This cross-validation method has been widely used in IDS researches, such as in [52], [53], and [54].

We also added the explanation on why we divide the data into 70%-30%% in second paragraph of sub-section IV.A, page 5.

The 70:30 data portion was used in [49]. The experimental results in [50] shows that the use of the 70:30 portion of training and testing data leads to the same level of accuracy as the portions of 80:20 and 60:40. Meanwhile, experimental result of using 70:30 data portion in other work [51] obtains high accuracy.

Reviewer#3, Concern # 3: There is not enough information about how the data set is processed.

Author response: Thank you for the comment. We realize that there was an error in the placement of information and we fix it.

Author action: We updated the manuscript by revising the first para in sub-section III.A as follows.

This study uses MachineLearningCSV data, which is part of the CICIDS-2017 dataset from ISCX Consortium. MachineLearningCSV consists of eight (8) traffic monitoring sessions, each in the form of a CSV file. This file contains normal traffic defined as “Benign” traffic and anomaly traffic called as “Attacks” traffic. The attack traffics are detailed more as in the second column of Table 2. Other than normal traffic and benign traffic, there are 14 types of attacks in this dataset.

And information about how the data is processed is explained in first para of sub-section IV.A in page

The eight CSV files as listed in Table 2 are combined into one CSV file. Next, to process the dataset using Weka software, this CSV file is converted into the ARFF file. The experiment uses only 20% of MachineLearningCSV data. There are 78 regular features and one class label used in this study. The dataset contains two features or columns named “Fwd Header Length” that make it as redundant features, so one of those columns must be removed. Thus, after removing the redundant features, only 77 features are available to be analyzed. As described in the CICIDS-2017 data prone to high-class imbalance will impact low detection accuracy and high false alarm. By adopting solution suggested by Karimi et al. [30] a new labeling attack traffic is introduced as listed in Table 3. The data type of the 77 features are numerical data type, so no data transformation is required to feed the data into Weka software.

Reviewer#3, Concern # 4: It is not mentioned what the parameters of the algorithms are selected according to.

Author response: Thank you for the concern, we agree to add explanation on the parameters to select the classifier algorithms.

Author action: We updated the manuscript by explaining the parameters as criteria in selecting the classifier algorithms in subsection III.C, para 1, page 4 as follows.

The main consideration on parameters for selecting classifier algorithms in this work is good performance in term of accuracy, learning ability, scalability, and speed. Having done some researches on several previous works that support the consideration, five algorithms are considered, they are: Random Forests, Bayesian Network, Random Trees, Naive Bayes and J48 classifiers to be experimented in this work. Research work by Hadi [20]) which states that random forest trees are strong learners and have good performance in detecting attacks based on the features resulted by Information Gain feature selection. Reazul et al. [39] reveals that the ability of Bayesian Network in classifying attacks outperforms other algorithms. According to Geetha and Kannan [57], Random Tree is an algorithm that has scalability and efficiency. Naive Bayes is a classification algorithm that is able to identify class labels faster than other algorithms because it has a low complexity of the model [55]. Sahu and Mehtre [15] conclude that J48 algorithm has good accuracy in classifying attacks.

Reviewer#3, Concern # 5: Total number of data should also be added to the Table II.

Author response: Thanks and agree with the suggestion

Author action: We updated the manuscript by adding the total number of instance/data in Table II, Table III and Table IV.

Reviewer#3, Concern # 6: accuracy in the result tables is shown as 99%. However, other results are in the form of 0.99. They should make it all in one format. It creates enormous complexity in terms of review.

Author response: Thanks for the sharp observations; we apologize for not being careful in presenting the data.

Author action: We revised the entries in Table VI to Table XXII for the consistency purpose.

Reviewer#4, Concern # 1: In Sub-Section IV. B, please specify whether your feature selection methodology is filter based, wrapper based or hybrid (a combination of both).

Author response: We use a filter based feature selection methodology

Author action: We updated Sub-section IV.B by mentioning explicitly that the experiments use filter-based FS technique. (Please refer to page 6, line 2-5) as follows:

As mentioned in sub-section 3.C, the feature selection in this experiment uses a filter-based approach. In other words, the feature selection filters throughout the weight scores, in which features are grouped based on the score of the feature's weight.

We also mentioned it in sub-section III.C, para 1, line 1-4, page 4, as follows:

Information Gain is the most used feature selection technique. It is a filter-based feature selection[28] , [30]. Information Gain uses a simple attribute rank and reduces noise that caused by irrelevant features then detects a feature that most have an information base, specific class.

Note: *References suggested by reviewers should only be added if it is relevant to the article and makes it more complete. Excessive cases of recommending non-relevant articles should be reported to ieeeaccesseic@ieee.org*