Network Intrusion Detection Leveraging Machine Learning and Feature Selection

Arshid Ali*, Shahtaj Shaukat†, Amreen Batool‡, Muazzam A Khan§, Jan Sher Khan¶, Arshad‖, Jawad Ahmad**
* Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan.
† Department of Electrical Engineering, HITEC University Taxila, Pakistan.
‡ Department of Computer Science, Tianjin Polytechnic University, China.
§ Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan.
¶ Department of Electrical and Electronics Engineering, University of Gaziantep, 27310 Gaziantep, Turkey.
‖ Institute for Energy and Environment, University of Strathclyde, United Kingdom.
** School of Computing, Edinburgh Napier University, United Kingdom.

***Abstract*—Handling superfluous and insignificant features in high-dimension data sets incidents led to a long-term demand for system anomaly detection. Ignoring such elements with spectral instruction not speeds up the analysis process but again facilitates classifiers to make accurate selections during attack perception stage, when wrestling with huge-scale and heterogeneous data. In this paper, for dimensionality reduction of data, we use Correlation-based Feature Selection (CFS) and Naïve Bayes (NB) classifier techniques. The proposed Intrusion Detection System (IDS) classifies attacks using a Multilayer Perceptron (MLP) and Instance-Based Learning algorithm (IBK). The accuracy of the introduced IDS is 99.87% and 99.82% with only 5 and 3 features out of 78 features for IBK. Other metrics such as precision, Recall, F-measure, and Receiver Operating Curve (ROC) also confirm the principal performance of IBK compared to MLP.**

***Index Terms*—Intrusion Detection System (IDS) , Correlation-Based Feature (CFS), Classifier subset evaluation, Multilayer Perceptron (MLP), Instance-Based Learning algorithm (IBK)**

I. Introduction

Over the past decades, automated data sharing has gained heightened concern with the rapid advancement of the Internet and computer technologies. For the high availability of Internet systems, anyone can use the Internet in a matter of seconds. Such a high availability and flexibility of access has raised many security issues [1]. Smart devices and services upon communication generate data, and there is a crucial need to secure this information from unwarranted access. To address this issue, we must design a smart IDS that can determine from the data and assure supervised access to the end-user system [2].

However, IDS learn from the data to observe the network traffic and analyze deviation from the predicted behavior of passing traffic. Based on the detection, we can classify IDS methods into two types: (a) Signature-based or Misuse Intrusion Detection System (M-IDS) and (b) Anomaly-based Intrusion Detection System (A-IDS) [3]. An IDS which identify attack by correlating the network traffic patterns with a pre-defined attack signature are Signature-based or Misuse Intrusion Detection System (M-IDS). The system which classifies the attack from the baseline behavior of the system is Anomaly-based Intrusion Detection Systems (A-IDS). An A-IDS learns from a normal attitude and pinpoints the attack upon intrusion.

To design an IDS applying ML techniques, the researcher has presented several means to enhance the efficiency of the system in terms of accuracy, precision, ROC, and computational time [4]. Different algorithms have been used to train the ML model [5]. Information Gain and Principle Component Analysis (IG-PCA) with ensemble classifier [6], Random Neural Network and an Artificial Bee Colony algorithm (RNN-ABC) [7], and numerous other ML techniques including Support Vector Machine (SVM), Multivariate Adaptive Regression Splines (MARS), and Artificial Neural Networks (ANN) are used to enhance the accuracy of the model [8]. In this paper, we aim to use dimensionality reduction methods to wipe out insignificant and superfluous features from data and then apply ML algorithms for intrusion detection.

The prime contributions of this paper are:

- We carry out the dimensionality reduction of the data set using CFS and Classifier subset evaluation (using NB as a classifier) method.
- To classify attacks, we use MLP and IBK classifiers.
- University of New Brunswick's (UNB), Canadian Institute for Cybersecurity (CIC) proposed CIC IDS-2017 data set [9]. The data set CIC IDS-2017 is used to educate and investigate the model. The CIC IDS-2017 data set contains benign and the most up-to date frequent attacks, which feature the normal real system data (PCAPs). PCAPs is an application programming interface (API) for catching system traffic.
- We compare IBK with MLP by modes of accuracy and other performance metrics, such as precision and build time.

We formulate the rest of the paper as follows. Section II reports the method, feature selection method and ML algorithms, likewise, section III reports results and consideration and section IV concludes the proposed method.

II. Research Methodology

In this paper, we use the CIC IDS-2017 [10] and investigate the port scan attack that occurs by sending packets to the

victim machine using a listening service to determine what ports are accessible on the victim machine.

Fig. 1 presents the proposed scheme comprising the following stage. For dimensionality reduction we use CFS and Classifier subset evaluation (NB as a classifier) to select an important feature from the data set. Likewise, MLP and IBK ML algorithms are used to develop and train the model.

We used Weka 3.8.4 [11] for data mining and ML. Weka comprises tools for data pre-processing, classification, regression, clustering and data visualization. It is farther effective-enhanced for building up different ML schemes. We test the proposed scheme in Weka 3.8.4, Intel 391 Core i5-380M (2.53 GHz) Central Processing Unit 392 (CPU) with 4 GB RAM.

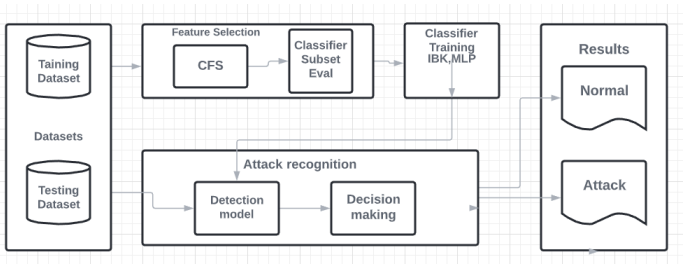

Fig. 1: The framework of the proposed model

A. Feature Selection

For dimensionality reduction, the feature selection method is used to eliminate extraneous and superfluous data from the data set. In data mining to interpret and envision the data we select relevant features from the data set through a feature selection method, that minimize the complication, build time of design, and storage needed for data set. However, there are many feature selection methods but, in this paper, we use the CFS and Classifier subset evaluation technique for features selection.

1) Correlation-based Feature Selection (CFS): CFS approach uses Best-First as a search method to pick out the most relevant features in the data set. To test the worth of a subset of attributes, CFS relate correlation along with the individual predictive ability of each feature between them. Using a correlation based heuristic evaluation function, CFS preferred the subsets of features having low intercorrelation while correlate with the class using the equation [12]:

$$Merit_s = \frac{k\overline{rfc}}{\sqrt{k + k(k-1)\overline{rff}}} \tag{1}$$

where $Merit_s$ represent the heuristic merit of attribute subsets s having k features, $\overline{rcf}$ is mean correlation between the class label s and each resource and $\overline{rff}$ is the mean correlation between two features [13].

2) Classifier Subset Evaluation (Naïve Bayes): Classifier subset evaluation also uses Best-First as a search method for feature selection. Classifier subset test attribute subsets on training data or a separate hold out the testing set. However, to evaluate the accuracy of a set of attributes classifier subset evaluation uses a classifier [14]. In this paper, we used NB classifier to evaluate the accuracy of a set of attributes. As NB is a statistical classifier, based on Bayes' theorem [15], specifies that all the features of data set are independent of each other.

B. Machine Learning Algorithm

ML algorithms are the key to designing an IDS. Different ML algorithms have different accuracy and build time. However, by implementing different ML algorithms we can observe different specifications of the design. In this paper we used MLP and IBK ML classifier, using CIC IDS 2017 data set file and implement these two ML classifiers using Weka explorer.

1) Multilayer Perceptron (MLP): MLP is the most used ML classifier, that is a Feed-forward artificial neural network, having one or more layers between the input and output layer of the network. Consider a network of different layers, the first layer will be as the input layer, and the last will be as output layer while the middle layers are the hidden layers of the network. The most typical neural model is feed-forward that maps an input data X to an output class Y using some function $f(x)$. Figure 2, illustrated the general structure of a feed-forward artificial neural network.

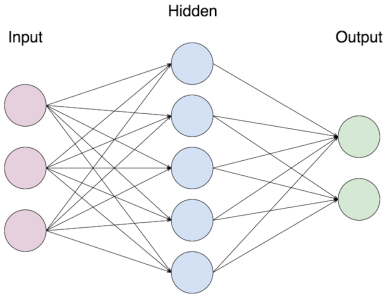

Fig. 2: Feed-forward artificial neural network

However, MLP mapped the input data into output data by adjusting the weight among its internal nodes. Using back propagation learning technique, a function $f(x)$

$f(x) : R^i \rightarrow R^o$ learns through MLP by training on data set [16], where $i, o \in Z^+$ represent the dimensional number of input i and output o, that can be calculated as [16]:

$$y = \phi(\sum_{i=1}^{n} w_i \mathbf{X} + b) \tag{2}$$

$$y = \phi(W^T \mathbf{X} + b) \tag{3}$$

Where ϕ is the activation function, w is the weights, X is the input data and b is the bias.

2) Instance-based learning algorithm (IBK): IBK can be both used for classification and regression. IBK can be also used for pattern recognition. IBK machine learning classifier belongs to lazy learning technique, also known as K Nearest-Neighbors classifier (K-NN). In such ML classifier, the raw training instances are used for prediction and it requires no learning for the training model. However, the basic concept of K-NN is that K-NN uses a majority poll between the k

most similar instances and the new instances, where the key factor is the distance to identify the similarity between the data vectors.

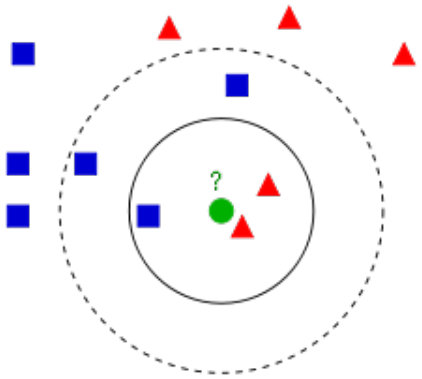

Fig. 3: Illustration of K-NN classification

IBK uses Euclidean distance for continuing data vectors and Hamming distance for discrete data vectors. Suppose we have pairs $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ where $x_i \in R^d$ and $y_i \in \{0, 1\}$ and Y is the class label of X;

To identify the k nearest for the new instance i, K-NN used the majority polls and weight distance between the data vectors to find the most similar data vector [17].

$$\left\{ \begin{array}{c} d^2\left(x_i, y_j\right) = \left\|x_1 - y_j\right\| \\ d^2\left(x_i, y_j\right) = \sum_{k=1}^{d}\left(x_i k - x_j k\right)^2 \end{array} \right\} \tag{4}$$

where $\left(x_i, x_j\right) \in R^d, x_i = \left(x_{i1}, x_{i2}, ...x_{id}\right)$.

III. Results and Discussion

In this section, we have extended the analyzed results of our proposed scheme and ML classifiers. We discussed in section II that feature selection is a key ingredient for data mining. However, we reduce the data set by applying the selected two feature selection method. The port scan data set of CIC IDS 2017 data set comprise 78 attributes, total instances 131860, and the sum of weights 131860. However, the data set has two class Benign and Attack (port scan) having 87899 and 43960 counts.

To figure out the efficiency of the proposed method, we conducted two experiments. According to the confusion matrix presented in Table I, these assessment metrics are used: Accuracy, Detection rate (DR, also recall or sensitivity), False Alarm Rate (FAR), F-measure, Precision, and ROC curve. The numerical calculations of the used evaluation metrics are explained in [18].

Where the True Positive (TP) is the number of actual attacks classified as attacks, True Negative (TN) is the number of normal instances classified as normal, False Negative (FN) is the number of attacks classified as normal instances, and False Positive (FP) is the number of normal instances classified as attacks [19], [20]. We demonstrate the computed confusion matrices in Table II, III, IV and V.

The attack forecast of the designed model is accurate only if it identifies attacks with a higher precision and low false-positive rate. Mathematically accuracy, precision and recall can be computed as [20].

TABLE I: Confusion matrix

Classified as		
	Normal	Attack
Normal	TP	FN
Attack	FP	TN

TABLE II: Confusion matrix of CFS MLP

Classified as		
	Normal	Attack
Normal	87897	2
Attack	519	43441

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{5}$$

Precision measures the exactness of a classifier.

$$Precision = \frac{(TP)}{(TP + FP)} \tag{6}$$

$$Recall = \frac{(TP)}{(TP + FN)} \tag{7}$$

By applying feature selection method in the WEKA environment. CFS method reduced the total number of attributes from 78 to 5 attributes. However, Classifier subset evaluation method by selecting NB as classifier reduced the total number of attributes to 3 attributes. The results of the feature selection show that both selection methods are good enough. However, the feature selection using Classifier subset evaluation (NB) reduced the total attributes to 3 and after applying ML algorithms, the features show high accuracy.

To evaluate these selected features from the data set, we apply two ML algorithms to data to build and train the model. We applied the ML algorithms to CFS selected feature and to classifier subset evaluation feature. However, the performance of ML algorithms is very effective to build time, accuracy, recall, and precision. We show the detailed results in Tables VI and VII.

Fig[2-5], shows detailed results and analysis. The highest accuracy got in the data set is 99.87% by IBK classifier using CFS features selection method. Likewise, the accuracy got by MLP classifier using the CFS features selection method is 99.67%. More in-depth, the time taken to build the model by IBK is 0.18 seconds while 55.46 seconds by MLP. We get the highest accuracy using classifier subset evaluation (NB as a classifier) is 99.82% by IBK classifier while 99.75% accuracy by MLP classifier using the same feature selection method. The total time taken to build a model by IBK is 0.18 seconds while 36.72 seconds by MLP classifier. By analyzing the feature selection methods, the results show that classifier subset evaluation using NB as a classifier is better than CFS selection method because the number of attributes reduced by classifier subset evaluation is more than CFS. However, both methods are well sufficient, and selected features show high accuracy.

In ML, two things are very prominent. The accuracy of the ML algorithm and the time taken to build the model. In

TABLE III: Confusion matrix of CFS IBK

	Classified as	
	Normal	Attack
Normal	87894	5
Attack	155	43805

TABLE IV: Confusion matrix of Classifier MLP

	Classified as	
	Normal	Attack
Normal	87797	102
Attack	226	43734

this paper, we also consider these two primary metrics for the study. However, by correlating these two ML algorithms, IBK performed better than MLP although both ML algorithms have effective accuracy and precision. The important factor is the total time taken to build the model (TTBM). Thus IBK has less time to build that reduce the computational complexity. So the preliminary analysis shows that IBK is an effective ML algorithm.

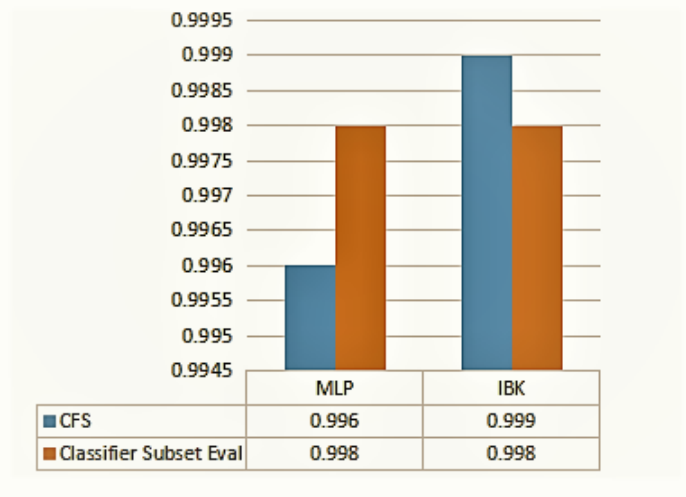

Fig. 4: Precision results.

TABLE V: Confusion matrix of classifier IBK

	Classified as	
	Normal	Attack
Normal	87899	0
Attack	226	43734

TABLE VI: Evaluation measures for CFS feature selection method

	ML Classifier	
Parameter	MLP	IBK
Accuracy	99.67	99.87
Precision	0.996	0.999
Recall	0.996	0.999
TR Rate	0.996	0.999
FR Rate	0.008	0.002
MMC	0.991	0.997
F Measure	0.996	0.999
ROC	0.998	1.000
Total time taken to build model (sec)	55.46	0.18

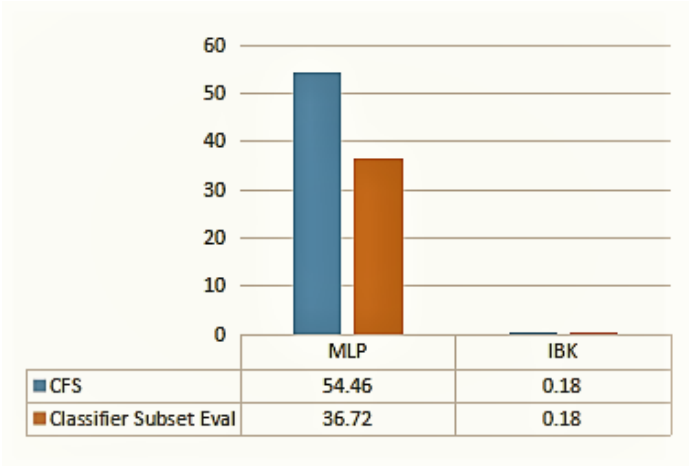

Fig. 5: Total time utilisation to build the model.

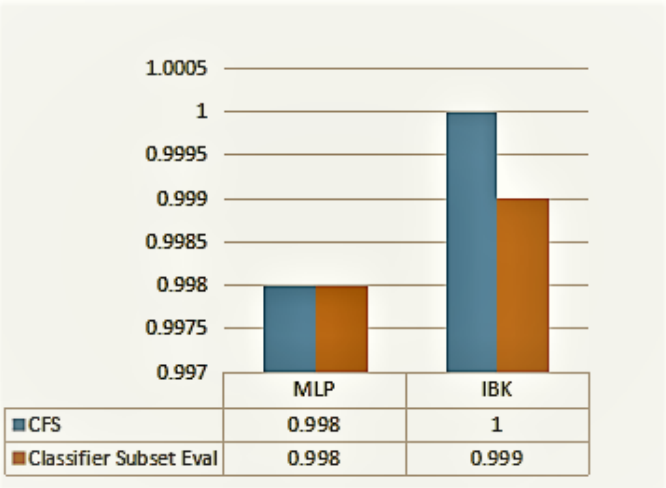

Fig. 6: Receiver Operating Characteristic results.

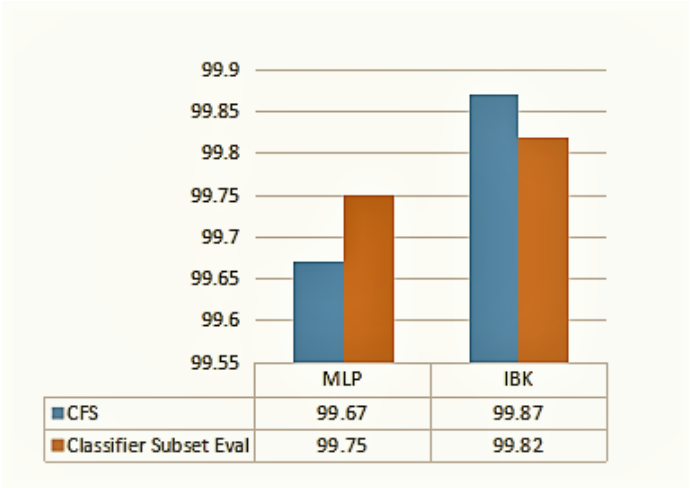

Fig. 7: Accuracy results.

TABLE VII: Evaluation measures for Classifier subset evaluation method

	ML Classifier	
Parameter	MLP	IBK
Accuracy	99.75	99.82
Precision	0.998	0.998
Recall	0.998	0.998
TR Rate	0.998	0.998
FR Rate	0.004	0.003
MMC	0.004	0.003
F Measure	0.994	0.998
ROC	0.998	0.999
Total time taken to build model (sec)	36.72	0.18

IV. Conclusion

With the continuous advancement of the intrusion detection system, many researchers proposed several approaches for the performance improvement of the system by using ML algorithms to secure the system. In this work, we compared the novel approach by features selection using a correlation-based feature selection and classifier subset evaluation method to select the relevant features from the data set, hence reducing complexity. In the proposed work, the CFS method reduced the total 78 attributes to 5 attributes, the classifier subset evaluation method reduced the total attributes to 3 attributes. Then MLP and IBK algorithms apply to a reduced number of features. Several parameters such as accuracy, precision, Recall, F-measure and ROC prove that IBK is more accurate than MLP.

References

[1] Larijani, H., Ahmad, J. and Mtetwa, N., 2019, July. A heuristic intrusion detection system for Internet-of-Things (IoT). In Intelligent computing-proceedings of the computing conference (pp. 86-98). Springer, Cham.

[2] Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G. and Vázquez, E., 2009. Anomaly-based network intrusion detection: Techniques, systems and challenges. computers security, 28(1-2), pp.18-28.

[3] Tang, Y. and Chen, S., 2007. An automated signature-based approach against polymorphic internet worms. IEEE Transactions on Parallel and Distributed Systems, 18(7), pp.879-892.

[4] Liao, H.J., Lin, C.H.R., Lin, Y.C. and Tung, K.Y., 2013. Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), pp.16-24.

[5] Sarnovsky, M. and Paralic, J., 2020. Hierarchical intrusion detection using machine learning and knowledge model. Symmetry, 12(2), p.203.

[6] Salo, F., Nassif, A.B. and Essex, A., 2019. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. Computer Networks, 148, pp.164-175.

[7] Qureshi, A.U.H., Larijani, H., Mtetwa, N., Javed, A. and Ahmad, J., 2019. RNN-ABC: A new swarm optimization based technique for anomaly detection. Computers, 8(3), p.59.

[8] Mukkamala, S., Sung, A.H. and Abraham, A., 2005. Intrusion detection using an ensemble of intelligent paradigms. Journal of network and computer applications, 28(2), pp.167-182.

[9] A. Yulianto, P. Sukarno, and N. A. Suwastika, "Improving adaboostbased intrusion detection system (ids) performance on cic ids 2017 dataset," in Journal of Physics: Conference Series, vol. 1192, p. 012018,IOP Publishing, 2019.

[10] A. Gharib, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Anevaluation framework for intrusion detection dataset," in 2016 International Conference on Information Science and Security (ICISS), pp.1–6,IEEE, 2016.

[11] G. Holmes, A. Donkin, I.H. Witten, Weka: a machine learning workbench, in:Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on, IEEE, 1994, pp. 357–361.

[12] Noviyanto, A., Isa, S.M., Wasito, I. and Arymurthy, A.M., 2011. Selecting features of single lead ECG signal for automatic sleep stages classification using correlation-based feature subset selection. IJCSI International Journal of Computer Science Issues, 8(1-5).

[13] Hall, M.A. and Smith, L.A., 1998. Practical feature subset selection for machine learning.

[14] Hao, H., Liu, C.L. and Sako, H., 2003, August. Comparison of genetic algorithm and sequential search methods for classifier subset selection. In Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. (pp. 765-769). IEEE.

[15] Koc, L., Mazzuchi, T.A. and Sarkani, S., 2012. A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier. Expert Systems with Applications, 39(18), pp.13492-13500.

[16] Leung, H. and Haykin, S., 1991. The complex backpropagation algorithm. IEEE Transactions on signal processing, 39(9), pp.2101-2104.

[17] Aha, D.W., Kibler, D. and Albert, M.K., 1991. Instance-based learning algorithms. Machine learning, 6(1), pp.37-66.

[18] S. Elhag, A. Fernández, A. Bawakid, S. Alshomrani, F. Herrera, On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on intrusion detection systems, Expert Syst. Appl. 42 (1) (2015) 193–202

[19] K. Kumar and J. S. Batth, "Network intrusion detection with feature selection techniques using machine-learning algorithms," International Journal of Computer Applications, vol. 150, no. 12, 2016.

[20] A. Verma and V. Ranga, "Machine learning based intrusion detection systems for iot applications," Wireless Personal Communications,vol. 111, no. 4, pp. 2287–2310, 2020.