Hindawi
Complexity
Volume 2020, Article ID 8694796, 7 pages
https://doi.org/10.1155/2020/8694796

Research Article

Deep Learning-Based Efficient Model Development for Phishing Detection Using Random Forest and BLSTM Classifiers

Shan Wang[1,2] **Sulaiman Khan**[3] **Chuyi Xu**[1] **Shah Nazir**[3] **and Abdul Hafeez**[4]

[1]*School of Information Engineering, East China Jiao Tong University, Nanchang 330013, China*
[2]*Ganzhou HPY Technology Co., Ltd., Ganzhou 341000, China*
[3]*Department of Computer Science, University of Swabi, Swabi, Pakistan*
[4]*Department of Computer Science, UET Jalozai, Jalozai, Pakistan*

Correspondence should be addressed to Shan Wang; patrick_shan@163.com

Received 25 July 2020; Revised 11 September 2020; Accepted 14 September 2020; Published 24 September 2020

Academic Editor: M. Irfan Uddin

With the increase in the number of electronic devices and developments in the communication system, security becomes one of the challenging issues. Users are interacting with each other through different heterogeneous devices such as smart sensors, actuators, and many other devices to process, monitor, and communicate different scenarios of real life. Such communication needs a secure medium through which users can communicate in a secure and reliable way so that their information may not be lost. The proposed study is an endeavor toward the detection of phishing by using random forest and BLSTM classifiers. The experimental results of the proposed study are promising in phishing detection, and the study reflects the applicability of the proposed algorithms in the information security. The experimental results show that the BLSTM-based phishing detection model is prominent in ensuring the network security by generating a recognition rate of 95.47% compared to the conventional RF-based model that generates a recognition rate of 87.53%. This high recognition rate for the BLSTM-based model reflects the applicability of the proposed model for phishing detection.

1. Introduction

The field of information security is considered to be a major part of the communication system. The information security is the protection of data and information from illegal access so that the unwanted user cannot access or alter the data or contents of the data. Security plays an important role in transmission of data from one source to another. Since the last few decades, the number of attacks on information has risen, and intruders are trying to capture important information for their own benefits and use. The information security of an organization is highly dependent on different types of information of the organization [1–4].

Most of the communication is made through Internet of Things (IoT) and with connectivity of devices to a network. The smart devices are connected to communicate, process, compute, and monitor diverse real-time scenarios. As with the developments of technology, security remains one of the major concerns for communication and interaction [4–13]. The attacks on security and information can drastically put the information and network into big loss. The industries such as Internet of Medical Things (IoMT) allow to reduce the unnecessary visits to the hospital and alleviate burden on the medical care system by providing connectivity over a secure network between medical experts and patients. By doing so, they can save a lot of time and money [14, 15]. This is the reason behind that the number of IoT devices in healthcare networks is rising since the last few years. According to the analysis report of Frost and Sullivan, the IoMT market was worth $22.5 billion in 2016; this figure is expected to touch $72.02 billion in 2021 [14]. IoMT is sharply increasing such that 60% of global health care organizations have adapted it, and by the end of 2020, it is estimated to increase by 27% [15].

The contribution of the proposed study is to detect phishing by using random forest and binary long short-term

memory (BLSTM) classifiers. The experimental results of the proposed study are promising in phishing detection, and the study reflects the applicability of the proposed algorithms in the information security. After validating the applicability of the proposed model for different phishing datasets, it was concluded that the BLSTM-based phishing detection model is prominent in ensuring the network security by generating a recognition rate of 95.47% compared to the conventional RF-based model that generates a recognition rate of 87.53%. This high recognition rate for the BLSTM-based model reflects the applicability of the proposed model for phishing detection.

The organization of the paper is as follows: Section 2 shows the background study and related work to the proposed work. Section 3 represents the system design and implementation of the proposed work. Section 4 shows the results and discussion of the proposed study. The paper is concluded in Section 5.

2. Background Study

This section of the paper presents the related work reported in the proposed field and the background information about the deep learning-based BLSTM model and the random forest. The following sections briefly show the details of the background study.

2.1. Deep Learning-Based BLSTM Model. The emergence of deep learning algorithms has revolutionized the research area by facilitating the researchers with automatic feature extraction capabilities [16]. These automatic feature extraction capabilities not only enable the researchers to get rid of hectic job of selecting the most significant feature extraction techniques relevant to a certain problem but also ensure high recognition rates compared to the conventional classification algorithms. Schuster and Paliwal [17] proposed the bidirectional recur-rent neural network (BRNN) after designing the RNN in a forward (from left to right) as well as in a backward direction (from right to left). This allowed maintaining long-range context information about the past and future using bidirectional long short-term memory (BLSTM) [18]. The combination of the attributes of BRNN and LSTM models is collectively known as BLSTM. Gu et al. [19] proposed the convolution neural network for intrusion identification purposes, while He et al. [20] proposed the LSTM-based classification model for the development of the trusted model for pervasive computing. Figure 1 shows the conventional model for the BLSTM architecture.

In the proposed research, a seven-layer BLSTM architecture is followed for the classification and recognition purposes, where layer 1 acts as an input layer that accepts the feature set, while layer seven acts as an output layer that generates output in the form of legitimate or phishing websites. The remaining five layers are the hidden layers that decided the output based on the feature set. The rectified linear unit (ReLU) is followed as an activation function in the proposed research work. Due to its (BLSTM) automatic feature extractor and high applicability to certain problems, it generates optimum results compared to the other conventional models such as random forest as shown in the results and discussion section of the paper.

2.1.1. Cross-Validation Method. Data classification is done using holdout methods: 70% for training and 73% for testing in this study.

2.1.2. Performance Evaluation Metrics. Accuracy, model execution time, and ROC-AUC have been used as performance evaluation metrics to evaluate the performance of the proposed BLSTM-based phishing detection model. After testing for varying training and test sets, time consumption, accuracy measures, false-positive rate, false-negative rate, true-positive rate, true-negative rate, precision, $f1$ score, and comparative results with random forest accuracy results, it was concluded that the proposed model outperforms for the phishing website detection.

2.2. Random Forest. Random forest was introduced by Brieman [21]. This classification technique is considered as one of the most recent and popular classification tools which achieves high-performance results for different classification problems [22–25]. RF is a combo training technique that erects multiple decision trees, where each tree subscribes with a single vote for the assignment of the most frequent class based on the input data. Ellis et al. [26] suggested the concept of RF classification tool for the prediction of energy usage and type of physical activity based on the wrist- or hip-based accelerometer device. Pal supposed the use of the RF classifier for land cover classification [27]. For this purpose, he selected the data of an area in the UK named Landsat Enhanced Thematic Mapper Plus (ETM+), and 7 different land covers were used. The accuracy results of the RF were compared with the SVM, and they were found more accurate. Dogru and Subasi [28] suggested the use of RF for traffic accident detection based on the perimeters of speed and distance calculated from the microscopic view based on the ad hoc network. Figure 2 represents a conventional model of the RF classifier for handwritten Pashto character recognition.

3. System Design and Implementation

This section of the paper presents the proposed methodology for the efficient detection of the phishing and legitimate website, the database followed for the simulation purposes, the classifiers followed for the identification purposes, and the results.

3.1. Proposed Model. Figure 3 represents the model of the proposed phishing detection system. The proposed research work improved the detection capabilities of the proposed model by developing a hybrid feature set from evaluating the overall thirty different feature sets. This hybrid feature vector consists of ten new features that promise in resulting with high accuracy rates and low computational costs for the

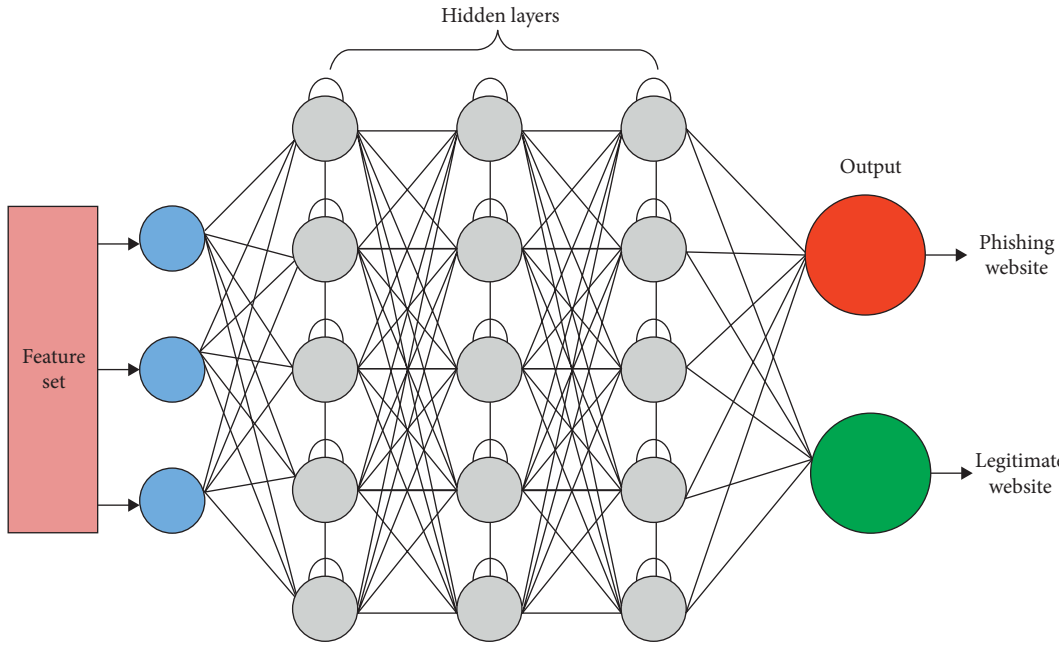

Figure 1: Generic BLSTM model.

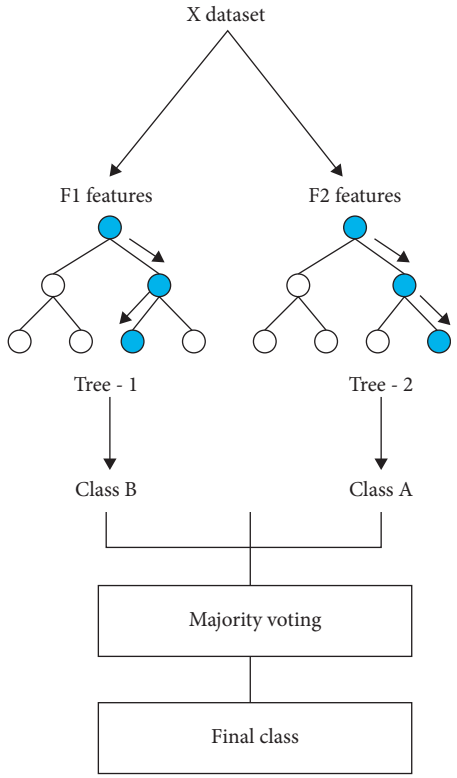

Figure 2: A conventional RF model for handwritten Pashto character recognition.

proposed model. This model works on by developing a hybrid feature vector by identifying the relationship between the web contents and the URL of the webpages. This feature vector is based on the hyperlink information of the webpages. The extraction of the hyperlink information from the webpages is depicted in Figure 4. The website hyperlink feature is extracted using the web crawler. After the feature extraction phase, the next step is to classify the legitimate and the phishing website. Based on this hybrid feature vector, two different classification algorithms, random forest (RF) and the binary long short-term memory (BLSTM) architectures, are followed for the classification purposes.

The hyperlink feature extraction is depicted in Figure 4. A hybrid feature vector is developed based on this model shown in Figure 4.

3.2. Dataset. This research work is implemented using the "phishing website dataset (https://archive.ics.uci.edu/ml/machine-learning-databases/00327/), access date: 15 July, 2020." This dataset accumulated 2456 instances and 30 different attributes.

3.3. Training with the Classifier. For the classification, the proposed model uses two different classification architectures named random forest and the binary long short-term memory. Since we have to classify two different classes, i.e., the legitimate and the phishing websites, it was more suitable for us to use the BLSTM model for the recognition process. The comparative results are performed with the random forest model to check the applicability of the proposed system. Both of these models are trained with the selected feature set.

After training the classifier, the proposed model is capable of deciding whether a new webpage is phishing or legitimate. For the training phase of the proposed model based on BLSTM architecture, different parameters are considered which are shown in Table 1.

The results of the proposed model based on BLSTM architecture are depicted in Figure 3. An overall accuracy of 95.47% is generated for the proposed model. This dataset has 30 different keywords and 2456 varying instances, so using

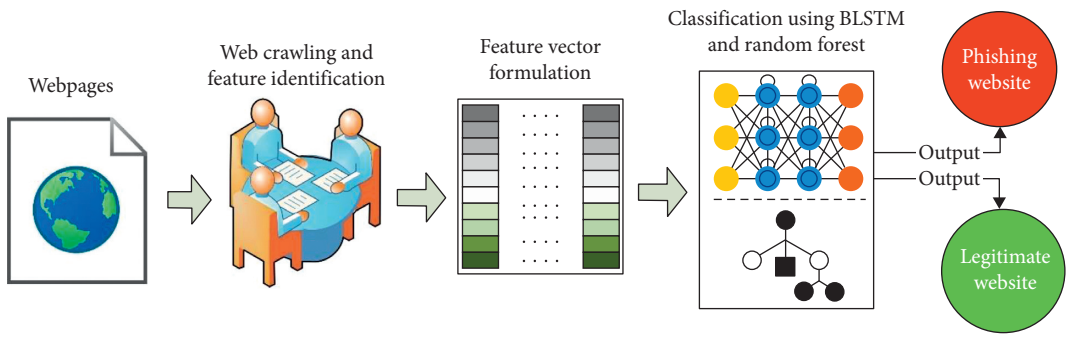

Figure 3: Proposed methodology.

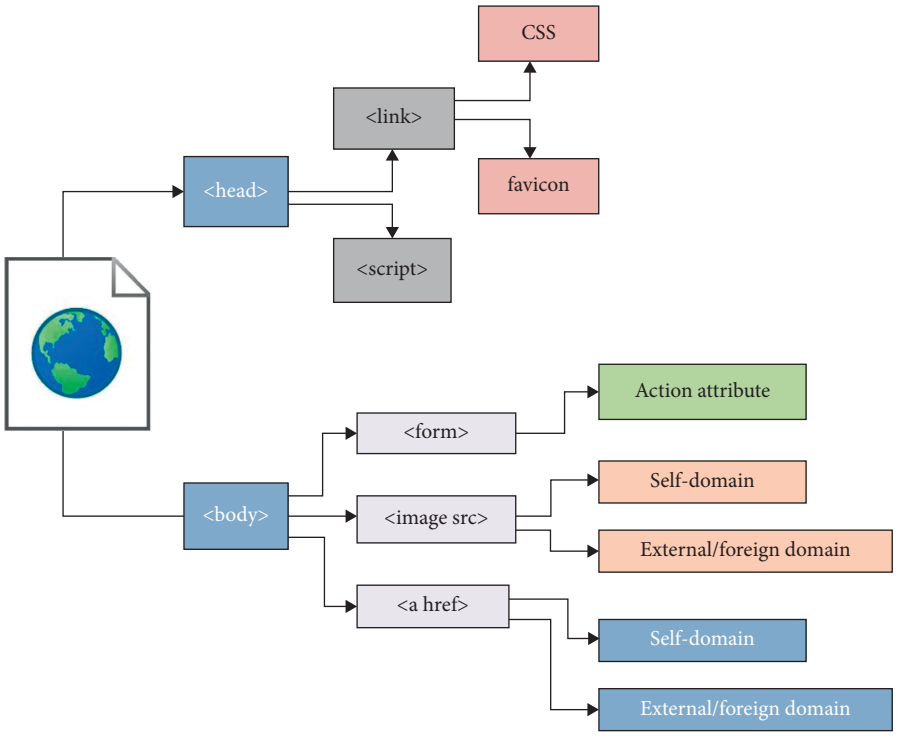

Figure 4: Hyperlink feature extraction process.

Table 1: LSTM model description.

S/no	Parameters	Description
1.	Total number of layers	7
2.	No. of hidden layers	5
3.	Activation function	ReLU
4.	Total number of instances	2456
5.	Number of total keywords	30

whole of the data for the training purpose takes a long time. To overcome this issue, only 10% of the data is used for the training purpose of the proposed model that contains 245 instances of the data. Additional performance evaluation metrics such as accuracy, processing time, and AUC have been computed for the model performance evaluation purpose. According to Table 2, *h4* achieved high performance as compared to other deep hidden layers' networks. Hidden layer 4 accuracy is 95.88%, AUC is 95.68%, and time consumption is 165 seconds.

The accuracy and ROC-AUC of different BLSTM networks have been graphically presented in Figure 5. These high-accuracy results show the applicability of the proposed model for the targeted problem.

The processing time of the proposed model based on BLSTM based on the number of hidden layers is depicted in Figure 6. From the figure, it is concluded that the proposed model performs very well, and within a limited time, it generates optimum accuracy results as depicted in Figure 5.

A confusion matrix is shown in Table 3 to check the performance of the proposed algorithm.

The output of the proposed model is evaluated after varying training and test sets. The phishing and the legitimate websites are decided based on the output values. If the output value generated is a value of zero, then it was reported as the phishing website; otherwise, a legitimate website. The applicability of the proposed model is also explored using the area under the ROC (receiver operating

Table 2: Training parameters of the BLSTM network.

BLSTM architecture	Hidden layers				
	$h1$	$h2$	$h3$	$h4$	$h5$
Training instances	245	245	245	245	245
Validating instances	100	100	100	100	100
Learning rate	0.001	0.0001	0.0001	0.01	0.0001
Activation function	ReLU	ReLU	ReLU	ReLU	ReLU
Number of epochs	200	600	800	900	1000
Training time (s)	130	150	160	165	200
Accuracy (%)	81.97	89.02	95.87	95.88	94.30
AUC (%)	81.27	88.52	93.47	95.68	94.53

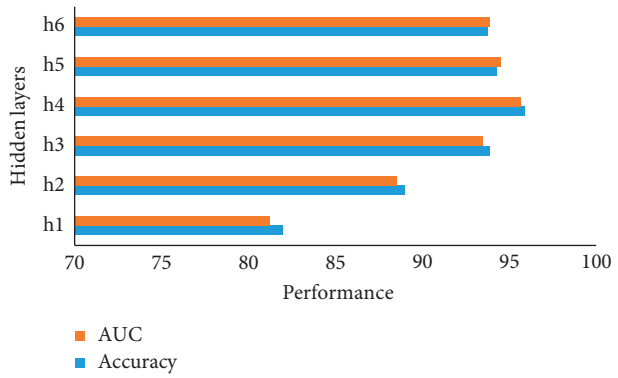

Figure 5: Proposed model results using BLSTM based on different hidden layers.

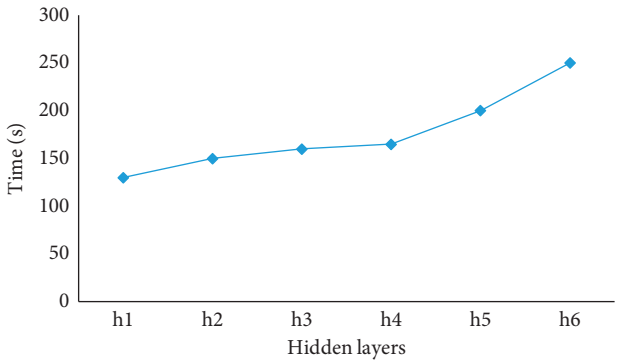

Figure 6: Time consumption for the proposed model using the BLSTM model.

Table 3: Confusion matrix.

True values		
	Legitimate (true)	Phishing (false)
Predicted values		
Legitimate (true)	True positive rate	False positive rate
Phishing (false)	False negative rate	True negative rate

characteristic) curve to find an optimal metric of precision. In the proposed experimental work, the area under the ROC curve for the phishing website is depicted in Figure 7, and it depicted that our model generates high-accuracy results for the classification of phishing and legitimate websites.

The results of the proposed model are also tested using different performance metrics such as accuracy, time consumption, precision, false positive rate, true positive rate, false negative rate, true negative rates, and $f1$ score. Table 4 represents the corresponding results of the proposed model based on the performance metrics.

4. Results and Discussion

The applicability of the proposed model is evaluated by testing its capabilities with the random forest model in deciding the legitimate and the phishing website detection. The performance results are depicted in Figure 8. From Figure 8, it is concluded that the proposed model performs very well compared to other traditional models. This high

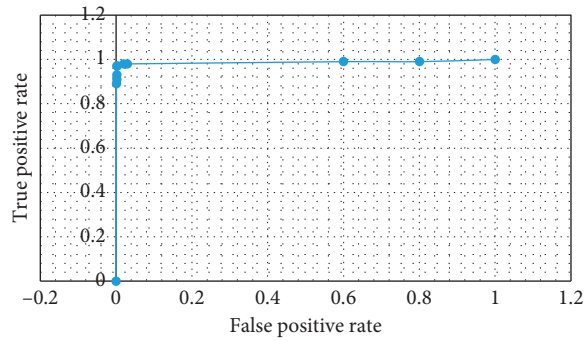

Figure 7: ROC curve for the proposed model.

Table 4: Performance results of the proposed model.

Accuracy (%)	*f*1 score (%)	Precision (%)	True positive rate (%)	True negative rate (%)	False positive rate (%)	False negative rate (%)
95.47	95.67	95.60	95.37	95.54	4.46	4.63

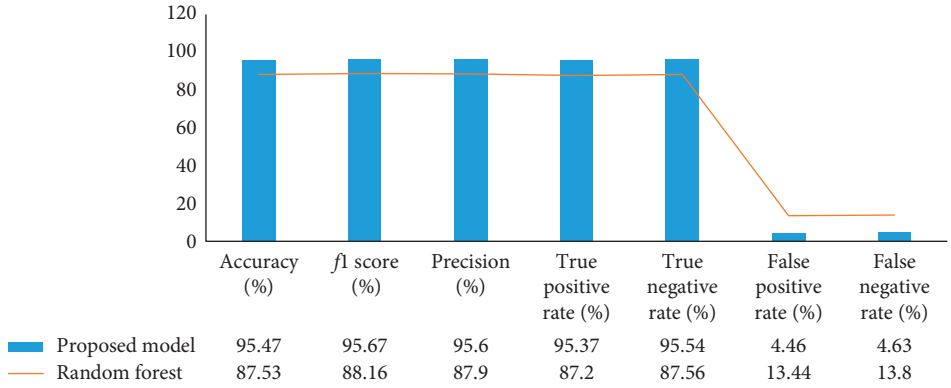

Figure 8: Performance comparison of the proposed model with the RF.

performance shows the applicability of the proposed model in deciding the phishing and legitimate website within a limited time and with high capabilities.

5. Conclusion

The security and privacy of the information security have been remaining as challenging concerns due to the heterogeneous nature of large-scale devices connected to the network and its vulnerability in the operating environment. During the transmission of data, it is likely that data can be handled maliciously and falsely by the hackers and intruders due to their depiction to attacks and vulnerabilities. Users are interacting with each other through different heterogeneous devices such as smart sensors, actuators, and many other devices to process, monitor, and communicate different scenarios of real life. Such communication needs a secure medium through which users can communicate in a secure and reliable way so that their information may not be lost. The proposed study is an endeavor toward the detection of phishing by using random forest and BLSTM classifiers. The experimental results show that the BLSTM-based phishing detection model is prominent in ensuring the network security by generating a recognition rate of 95.47% compared to the conventional RF-based model that generates a recognition rate of 87.53%. This high recognition rate for the BLSTM-based model reflects the applicability of the proposed model for phishing detection.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was sponsored in part by the National Natural Science Foundation of China (41965007).

References

[1] M. Li, S. Nazir, H. U. Khan, S. Shahzad, and R. Amin, "Modelling features-based birthmarks for security of end-to-end communication system," *Security and Communication Networks*, vol. 22, 2020.

[2] H. U. Rahman, A. U. Rehman, S. Nazir, I. U. Rehman, and N. Uddin, "Privacy and security—limits of personal information to minimize loss of privacy," in *Proceedings of the Presented at the Future of Information and Communication Conference*, New York, NY, USA, 2019.

[3] S. Nazir, S. Shahzad, S. Mahfooz, and M. N. Jan, "Fuzzy logic based decision support system for component security evaluation," *International Arab Journal of Information and Technology*, vol. 15, pp. 1–9, 2015.

[4] S. Nazir, S. Shahzad, M. Nazir, and H. U. Rehman, "Evaluating security of software components using analytic network process," in *Proceedings of the 11th International Conference on Frontiers of Information Technology (FIT)*, IEEE, Islamabad, Pakistan, pp. 183–188, 2013.

[5] H. H. Song, "Testing and evaluation system for cloud computing information security products," in *Proceedings of the Presented at the 3rd International Conference on Mechatronics and Intelligent Robotics (ICMIR-2019)*, Islamabad, Pakistan, 2020.

[6] B. K. Mohanta, D. Jena, U. Satapathy, and S. Patnaik, "Survey on IoT security: challenges and solution using machine learning, artificial intelligence and blockchain technology," *Internet of Things*, vol. 11, 2020.

[7] J. Yuan and X. Luo, "Regional energy security performance evaluation in China using MTGS and SPA-TOPSIS," *Science of the Total Environment*, vol. 11, pp. 1–11, 2019.

[8] X. Wu, S. Liu, Y. Sun, Y. An, S. Dong, and G. Liu, "Ecological security evaluation based on entropy matter-element model: a case study of Kunming city, southwest China," *Ecological Indicators*, vol. 102, pp. 469–478, 2019.

[9] X. Wang, J. Li, X. Kuang, Y.-A. Tan, and J. Li, "The security of machine learning in an adversarial setting: a survey," *Journal of Parallel and Distributed Computing*, vol. 130, pp. 12–23, 2019.

[10] M. Marwan, A. Kartit, and H. Ouahmane, "Security enhancement in healthcare cloud using machine learning," *Procedia Computer Science*, vol. 127, pp. 388–397, 2018.

[11] Z. Zhang, J. Wen, X. Wang, and C. Zhao, "A novel crowd evaluation method for security and trustworthiness of online social networks platforms based on signaling theory," *Journal of Computational Science*, vol. 12, 2017.

[12] Y. Cherdantseva, J. Hilton, O. Rana, and W. Ivins, "A multifaceted evaluation of the reference model of information assurance & security," *Computers & Security*, vol. 63, pp. 45–66, 2016.

[13] M. Jouini, L. B. A. Rabai, and R. Khedri, "A multidimensional approach towards a quantitative assessment of security threats," in *Proceedings of the Procedia Computer Science the 6th International Conference on Ambient Systems*, Networks and Technologies, New York, NY, USA, pp. 507–514, 2015.

[14] A.o.A.B Engineering. (Frost & Sullivan), "Internet of medical things revolutionizing healthcare," 2019.

[15] Fuzon, "Internet of Medical Things (IoMT): new era in healthcare industry," 2019.

[16] S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez, "KNN and ANN-based recognition of handwritten Pashto letters using zoning features," *International Journal of Advanced Computer Science and Applications*, vol. 9, pp. 570–577, 2018.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[18] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[19] Z. Gu, S. Nazir, C. Hong, and S. Khan, "Convolution neural network-based higher accurate intrusion identification system for the network security and communication," *Security and Communication Networks*, vol. 2020, 2020.

[20] Y. He, S. Nazir, B. Nie, S. Khan, and J. Zhang, "Developing an efficient deep learning-based trusted model for pervasive computing using an LSTM-based classification model," *Complexity*, vol. 2020, p. 4579495, 2020.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] S. Bernard, S. Adam, and L. Heutte, "Using random forests for handwritten digit recognition," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition ICDAR 2007*, pp. 1043–1047, New York, NY, USA, 2007.

[23] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*, Springer Science & Business Media, Berlin, Germany, 2013.

[24] J. Greenhalgh and M. Mirmehdi, "Traffic sign recognition using MSER and random forests," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012*, pp. 1935–1939, New York, NY, USA, 2012.

[25] N. J. Tustison, K. L. Shrinidhi, M. Wintermark et al., "Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTsR," *Neuroinformatics*, vol. 13, no. 2, pp. 209–225, 2015.

[26] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers," *Physiological Measurement*, vol. 35, no. 11, p. 2191, 2014.

[27] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[28] N. Dogru and A. Subasi, "Traffic accident detection using random forest classifier," in *Proceedings of the Learning and Technology Conference (L&T)*, pp. 40–45, New York, NY, USA, 2018.