

Received 2 July 2022, accepted 19 July 2022, date of publication 28 July 2022, date of current version 3 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3194672

RESEARCH ARTICLE

Multilayer Stacked Ensemble Learning Model to Detect Phishing Websites

LAKSHMANA RAO KALABARIGE¹, ROUTHU SRINIVASA RAO², AJITH ABRAHAM³,
AND LUBNA ABDELKAREIM GABRALLA⁴

¹AI Research Laboratory, Department of Computer Science and Engineering, GMRIT, Rajam 532127, India

²Department of CSE, GITAM School of Technology, GITAM (Deemed to be University), Visakhapatnam 530045, India

³Machine Intelligence Research Labs, Auburn, WA 98071, USA

⁴Department of Computer Science and Information Technology, College of Applied, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Routhu Srinivasa Rao (srouthu@gitam.edu)


This work was supported by the Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, through Researchers Supporting Project PNURSP2022R178.

ABSTRACT Phishing is a cyber attack that tricks the online users into revealing sensitive information with a fake website imitating a legitimate website. The attackers with stolen credentials not only use them for the targeted website but also can be used for accessing the other popular legitimate websites. There exists many anti-phishing techniques, toolbars, extensions to counter the phishing sites but still the phishing attacks are major concern in the current digital world. In this paper, we propose a multilayered stacked ensemble learning technique which consists of estimators at different layers where the predictions of estimators from current layer are fed as input to the next layer. From the experimental results, it is observed that the proposed model achieved a significant performance when evaluated with different datasets with an accuracy of ranging from 96.79% to 98.90%. The proposed model is evaluated with datasets from UCI(D1), Mendeley 2018(D2) and Mendeley 2020(D3,D4). The proposed model achieved detection rate of 97.76% with D1 dataset, achieved an accuracy of 98.9% with D2 dataset. Finally, the technique is tested with D3 and D4 which resulted in accuracy of 96.79% and 98.43% respectively. Also, the proposed model outperformed baseline models corresponding to datasets with a significant difference in accuracy and F score metrics.

INDEX TERMS Phishing, anti-phishing, meta learner, ensemble, stacking, machine learning.

I. INTRODUCTION

Now a days, the ease in communication over internet [1] has brought revolutionary changes. This positive transformation drastically increases the number of internet users. At the same time adversaries make use of this opportunity to steal sensitive credentials of an internet user by creating phishing websites or sending fake emails to naive online users. Phishing is an online attack in which phisher sends a fake email or replica of original website to all the online users as bait and waits for the innocent users to fall as prey. According to the phishing survey conducted by Anti-Phishing Working Group(APWG),¹ it is observed that the total number of phishing websites are increased in the year 2020 as shown

The associate editor coordinating the review of this manuscript and approving it for publication was Li He .

¹<https://apwg.org/trendsreports/>

in figure 1 when compared with all the four quarters of 2019. Similarly, there are 165,772 phishing sites in first quarter of 2020 and it was slightly increased from 162,155 detected phishing sites in quarter four of 2019. In general, phishing attack can be designed in two ways such as

- Sending fake e-mail: a spoofed e-mail sent to the users in the name of legitimate company or organization
- Sending Replica of website: attacker creates and launches replica of original website on Twitter, Facebook, google and on other social media platforms. These phishing websites also uses the green padlock and Hypertext Transfer Protocol Secure (HTTPS) to make the users to believe that this phishing site as legitimate one.

Many methods were proposed in the literature to detect and prevent phishing. Some of these techniques are summarized as follows.

- Blacklist (Rao and Pais [2], Whittaker *et al.* [3], Rao and Pais [4], Ma *et al.* [5]): The database of phishing URLs is known as blacklist. These URLs are blocked by the latest browsers such as Chrome, Opera, Mozilla, etc. However, this technique fails in preventing zero-day phishing sites.
- Feature extraction (Rao and Pais *et al.* [6], Chou *et al.* [7], Shahriar and Zulkernine *et al.* [8]; Rao and Pais [9]): The features are extracted from different phishing websites and they are used to detect and prevent phishing attacks. However, the extracted features may not be available in all phishing sites. Hence, the feature extraction methods may not guarantee for the detection and prevention of all websites.
- Machine Learning (ML) (Ramana *et al.* [10], Rao *et al.* [11], Xiang *et al.* [12]): The blacklist and feature extraction techniques are not up to the mark in detecting and preventing phishing attacks. In this connection, classification models Rao and Pais [13] such as Decision Tree (DT), Random forest (RF), etc. are used in the detection process. From the existing literature (Rao and Pais [13], Whittaker *et al.* [3], Khonji *et al.* [14]), it is observed that the machine learning based methods could achieve 99% of accuracy in detecting phishing websites.

It is known that, the performance of ML algorithms depends upon quantity of training data and quality of extracted features from phishing websites. The traditional ML models are unable to capture multiple characteristics of data due to data diversity. Where as, ensemble learning is enough capable to extract diversified features, combines predictive results produced by multiple learning algorithms, and finally, achieves better predictive performance results via ensemble methods such as voting, stacking, blending, averaging, etc.

Hence, this piece of work proposes a Multi layer stacked ensemble model to detect and prevent phishing websites. The main intention of using stacked ensemble is due to the feature of harnessing the capabilities of range of well performed models in the task of classification. The proposed model is applied on two variants (small and large) of Mendeley Phishing Dataset (MPD).² The small variant phishing dataset is named as dataset_small and large variant named as dataset_full. Each dataset consists of 111 features with different number of instances. The contributions of our proposed model are given below.

- We proposed a multi layer staked ensemble model combining various classifiers at different layers for the detection of phishing sites.
- We conducted experiments on various datasets of different size (11 K, 10K, 58 K, 88 K) and different feature space (30,48,111,111) to evaluate the behavior of the model with varying dataset.

The remaining part of this paper is organized as follows. The section II describes literature of various ML based phishing detection and prevention techniques. Section III

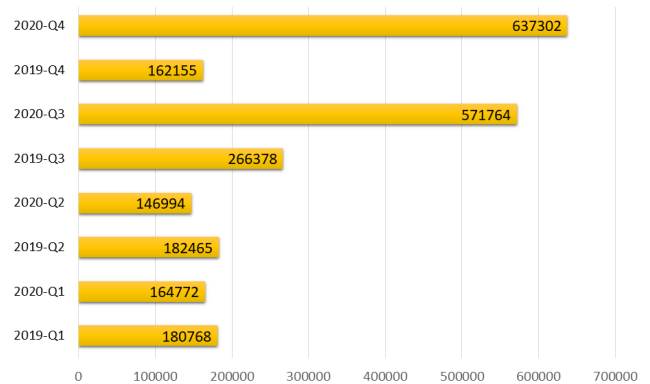


FIGURE 1. Unique phishing websites in each quarter.

describes the architecture and working of proposed model. The implementation of different phases of proposed model such as input dataset, data balancing, and Multi layer stacked ensemble model is given in Section III. The experimental results with various classifiers and baseline models are given in Section IV. The justifications, group of all findings and limitations of the proposed model is given in discussion section V. Finally, section VI concludes the proposed piece of work.

II. RELATED WORK

There exists many antiphishing techniques that use whitelists, blacklists, heuristics, visual descriptors, third party services and machine learning algorithms for the detection of phishing sites. Also, the techniques use the source code or URL for the feature extraction to classify the phishing sites. Out of all the antiphishing techniques, machine learning based techniques played a vital role in detection of phishing sites. Moreover, the existing literature demonstrates that machine learning based solutions could achieve performance of atleast 99%. Hence, we also attempted to use machine learning approach for the phishing detection. Some of the recent and popular antiphishing techniques are given below.

Shahrivari *et al.* [15] applied machine learning methods such as Logistic Regression (LR), Ada boost (AB), RF, K-Nearest Neighbor(KNN), Neural Networks, Support Vector Machine (SVM), Gradient boosting(GB), and XGBoost(XGB) on UCI phishing dataset. These models are evaluated based on accuracy, precision, recall, F1 score, training time, and testing time.

Priya *et al.* [16] uses Gravitational Search Algorithm (GSA) on UCI phishing dataset [17] to find optimal feature subset and applies classification algorithms such as RF, DT, Artificial Neural Networks (ANN), and KNN to classify phishing and legitimate websites.

Adi *et al.* [18] employs feature selection methods such as Information gain (IG), gain ratio, chi-square, and correlation-based feature selection on UCI phishing [17] dataset. The classification algorithms namely Naive-Bayes(NB), KNN, SVM, DT and ID3 are applied on the

²<https://data.mendeley.com/datasets/72ptz43s9v/1>

selected feature subset and observed decrease in accuracy of all these classification methods when compared with all 30 features. Similarly, Almseidin *et al.* [19] employs classification algorithms such as J48, RF, and Multi Layer Perceptron (MLP) along with feature selection methods information gain and ReliefF on mendely [20] dataset.

Babagoli *et al.* [21] applied harmony search (HS) and SVM after feature selection for the detection of phishing sites. The results show that the HS with SVM performs better than SVM alone. Similarly, Javadi-Moghaddam and Golami [22] applied feature selection method ReliefF followed by DT, NB, NN, KNN and SVM on UCI phishing [17] repository. Alotaibi and Alotaibi [23] proposed a phishing detection technique based on ensemble classifiers such as AdaBoost and XGBoost. These Ensemble methods tested on both UCI Phishing [17] and Mendely [20].

Zhu *et al.* [24] proposed Optimal Feature Selection based Neural Network (OFS-NN). The NN applied on the feature subset selected by OFS which selects optimal features based on threshold value set to each feature. The calculation of effective value of each feature determines its threshold. The proposed OFS-NN Zhu *et al.* [25] is an improvement to the Zhu *et al.* [24] and this model integrates Feature validity value (FVV) to each feature and selects features based on FVV.

Sharma *et al.* [26] applies classification algorithms such as DT, KNN, LR, MLP, NB, RF, SVM, and Stochastic Gradient Descent (SGD) on UCI [17] and Mendely [20] phishing data sets. The F1_Score of each classification model with all the features of each dataset is compared with the best features selected through chi-squared test, Keiser-Meyer-Olkin (KMO) test and Pearson Correlation test.

Singh and Tiwari [27] checks the performance of machine learning classification algorithms such as SVM, KNN, DT, RF, and Ensemble methods with various dimensionality reduction and feature selection methods. Gandotra and Gupta [28] applied eight machine learning algorithms such as Instance based learning (IB1), NB, J48, AdaBoost, DT, RF, and sequential minimal optimization (SMO) on UCI phishing dataset to classify legitimate and phishing websites. The accuracy of these classification models evaluated with all 30 features and compared with the best feature subset selected through feature selection methods. The results shows that the accuracy difference between all 30 and top 15 features is minute.

From the literature [29], it is observed that the performance of traditional machine learning models is unsatisfactory when it deals with a data which may be having large number of features, noisy data, etc. At the same time, the performance of the classifier or regressor may differ from one data to another. That is, it shows accurate results on one dataset and low accurate on other.

In this connection, researchers focused on ensemble learning methods to improve model performance by mitigating problems with data. The traditional ML models unable to capture multiple characteristics of data due to the diversity

in data. Where as, ensemble learning is enough capable to extract diversified features, combines predictive results produced by multiple learning algorithms, and finally, achieves better predictive performance results via ensemble methods such as voting, stacking, blending, averaging, etc. Hence, ensemble learning methods applied on phishing datasets [17], [20], and [30] achieve significant results comparatively with traditional learning algorithms.

A. ENSEMBLE MODELS TO DETECT PHISHING

The below are some of the ensemble learning models used in different domains to achieve high accuracy. The Mohammed AL-Sarem *et al.* [31] proposed Technique for Order Preferences by Similarity to Ideal Solution (TOPSIS) for instance based Arabic language authorship. The Xiaoxu Niu *et al.* [32] proposed ensemble empirical mode decomposition (EEMD) to predict displacement of landslide. The Ke Yan *et al.* [33] uses ensemble model to forecast ground surface settlement during tunnel construction to prevent serious damages due to landscape collapse. The Panagiotis Pintelas *et al.* [34] provided a state-of-art on various ensemble learning models.

The below are the ensemble learning models found in the literature to detect phishing websites. Al-Sarem *et al.* [35] proposed an optimized stacking ensemble method to detect phishing website. The genetic algorithm (GA) is used to achieve optimization by obtaining optimal parameters of ensemble learning algorithms such as Random forests, AdaBoost, XGBoost, Bagging, GradientBoost, and LightGBM. These ensemble methods applied on three phishing datasets such as UCI Phishing [17], Mendely [20] with 48 features and mendely-small variant [30] consists 58645 instances with 111 features, 27998 legitimate and 30647 phishing websites. This model archives 97.16%, 98.58%, and 97.35% accuracy respectively on these three datasets.

Basit *et al.* [36] proposed an ensemble model by integrating multiple classifiers. In which, multiple learning methods such as ANN, KNN, and Decision Tree (C4.5) are combined with an ensemble method namely Random Forest Classifier. That is, the proposed ensemble model uses RF as base classifier, implements in three combinations of ensemble methods such as RF with ANN, RF with C4.5 and RF with KNN and voting algorithm is applied on each combination. All three combinations applied on UCI phishing [17] dataset with batch size of 100, 10-fold cross-validation and evaluates each combination on four metrics such as Precision, Recall, F-measure and Accuracy. The experiment results shows that the RFC with KNN outperforms remaining two combinations by achieving 97.33% accuracy.

Tama and Rhee [37] compares the performance of ensemble classifiers (RF, Rotation Forest (RoF), GB, and XGB) against single classifiers (DT, regression tree, and credal decision tree) on a UCI phishing dataset.

These classifiers evaluated using Area under ROC curve (AUC) value with respect to different data splitting

procedures such as k-fold cross validation, subsampling, and bootstrap for training and testing of each single and ensemble learning algorithms. The experimental results shows that the RF followed by XGBoost algorithms performs better than rotation forest, GDM, and binary classifiers. Vaitkevicius and Marcinkevicius [38] proposed a stacked ensemble learning model to detect phishing websites. The proposed model stacks seven pre-trained models on mendely [20] dataset. This ensemble model stacks seven different models such as recurrent neural networks (RNN), Long Short-Term Memory network with Peepholes (LSTM-P), Long Short-Term Memory network (LSTM), Two convolution neural network (CNN)- Gated Recurrent Units (GRU), Gated Recurrent Units (GRU), and CNN-LSTM as five different ensemble models. The accuracy of these five ensemble models compared with Gradient Tree Boosting, AdaBoost, Random Forest, Multilayer Perceptron, Classification and Regression Trees, Support Vector Machine, GRU, CNN-GRU, LSTM, LSTM-P, Naive-Bayes, and simple RNN. From results, it is observed that the Gradient Tree Boosting, Ensemble-1, Adaboost, Ensemble-2, and Ensemble-3 models performs well with an accuracy of 97.42%, 97.30%, 97.28%, 97.25%, and 97.21%. it is observed that the Gradient Tree Boosting, Ensemble-1 models outperforms all other models.

Nagaraj *et al.* [39] proposed three twofold ensemble learning models such as Random Forest Neural Network model (RF_NN), Bagging Neural Network ensemble model (Bagging_NN) and Boosting Neural Network ensemble model (Boosting_NN). These three models are applied on UCI Phishing [17] dataset. In RF_NN, the predictions of RF are fed to a feedforward neural network in RF_NN, The predictions of bagging algorithm fed to a feedforward neural network in Bagging_NN, and the prediction of boosting algorithm given as input to the feedforward neural network. The experimental results shows that the RF_NN obtains 93.41% and outperforms RF, Bagging, Boosting, Bagging_NN, and Boosting_NN.

The Ensemble of KNN and Random committee using voting (EKRV) is proposed by Niranjan *et al.* [40] to detect the phishing sites. The proposed work consists of two phases such as pre-processing and classification. The 23 important features extracted from the UCI phishing [17] dataset in pre-processing phase. The experimental results shows that the voting method performs better than stacking. That is voting gives 97.4% of accuracy with 0.028 false-positive rate and stacking gives 97.1% of accuracy with 0.031 false-positive rate.

Taha [41] proposed an intelligent ensemble learning model to detect phishing websites. This piece of work combines outcome of six heterogeneous machine-learning models such as Random Forest, AdaBoost, XGBoost, Bagging, Gradient-Boost, and LightGBM to decide whether a website is phishing or legitimate. The results shows that the proposed approach achieves 95% of classification accuracy on UCI phishing dataset [17].

Zamir *et al.* [42] proposes two stacking approaches such as Stacking1 which combines NN, RF and bagging models and Stacking which combines KNN, RF and bagging models. These models evaluated on UCI phishing [17] dataset. The stacking1 achieves 97.4% and stacking2 achieves 97.2% of accuracy.

Ubing *et al.* [43] proposed a majority voting based ensemble learning approach, which considers Gaussian naive Bayes, SVM, KNN, LR, multilayer perceptron NN, GB and RF classifiers. In majority voting technique the outcome of maximum number of classifiers will be treated as final outcome. This model employs UCI phishing [17] dataset for evaluation and achieves 95.5% accuracy.

Adeyemo *et al.* [44] proposed an ensemble-based Logistic Model Trees (LMT) to detect phishing websites. This approach combines Logistic Regression with two different tree induction methods such as Adaboost(AB) and Bagging(BG). The combination of AB and LR is named as AdaBoostLMT. The BGLMT is a combination of BG and LR. This model obtains 97.18% of accuracy in detecting phishing websites.

Subasi and Kremic [45] used AdaBoost and MultiBoost with a combination of six diverse base classifiers such as KNN,ANN,SVM,DT,RF, and RoF. From the results, it is observed that the AdaBoost with SVM obtains 97.61% and MultiBoosting with RoF achieves 97.30% accuracy.

Al-Mekhlafi *et al.* [46] proposed optimized stacking ensemble model. the optimization of parameters achieved with Genetic Algorithm(GA). This model applies different ensemble models such as RF, AB,XGB,BA,GB,and LightGBM with and without GA. Form the results it is observed that the GA with GB, GA with XGB and GA with BA performs well on UCI [17] dataset.

III. PROPOSED MODEL

The proposed model includes the idea of layer wise Stacked Ensemble Learning for the phishing detection. The stacked ensemble learning builds layers where each layer encompasses required number of estimators e_1, e_2, \dots, e_n and finally, stacks all the layers. The multi layer stacked ensemble consists two phases, layers with learners/estimators and meta-learner. The architecture of the proposed model is shown in figure 3 and its working is as follows:

- 1) First it initializes the Model.
- 2) Develops required number of layers. Where, each layer encompasses required number of estimators/learners and each layer should be added to the Model.
- 3) Add meta-learner as last layer.
- 4) Finally, model training and prediction takes place.

In stacked ensemble learning, output of one layer passed as input to the next layer as shown in figure 3. The estimators such as Random Forest, Logistic Regression, K Nearest Neighbors etc denoted as e_1, e_2, \dots, e_n can be either used for the classification or regression. The estimators within the

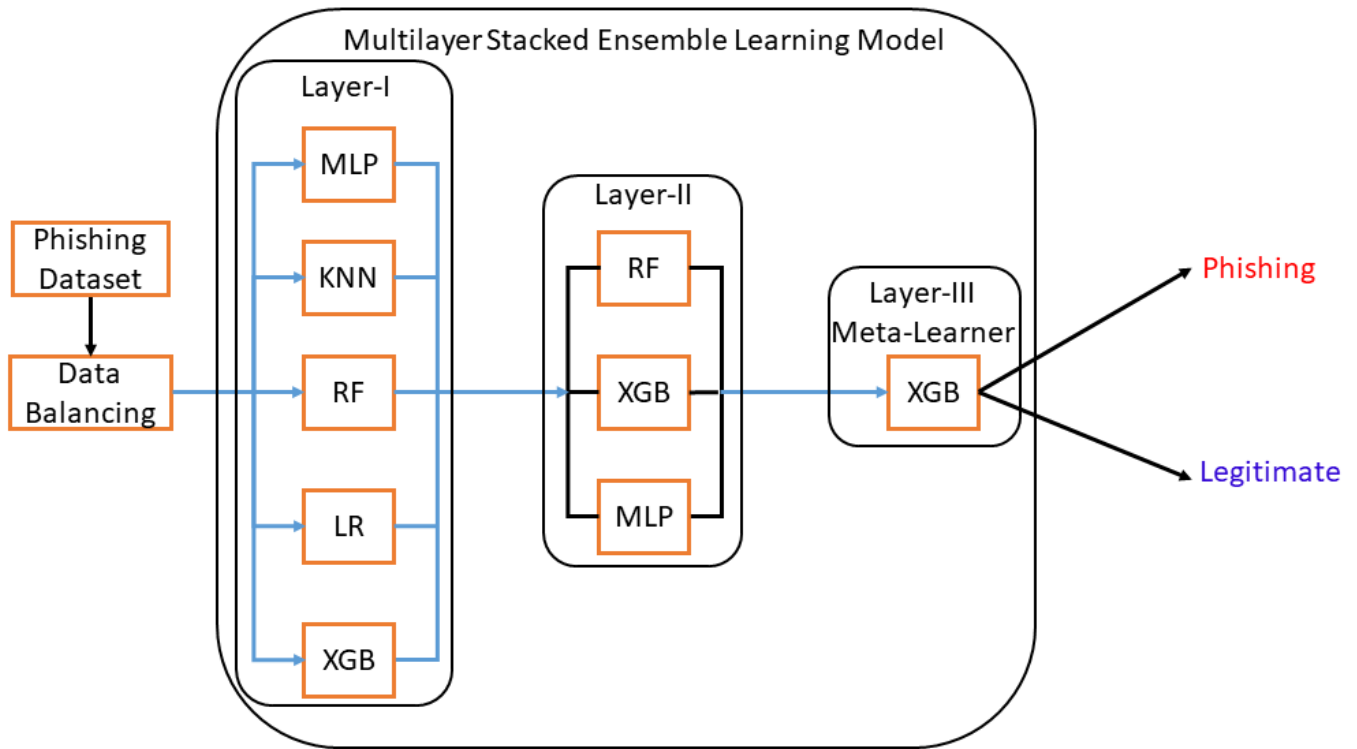


FIGURE 2. Proposed multi layer stacked ensemble learning model(MLSELM).

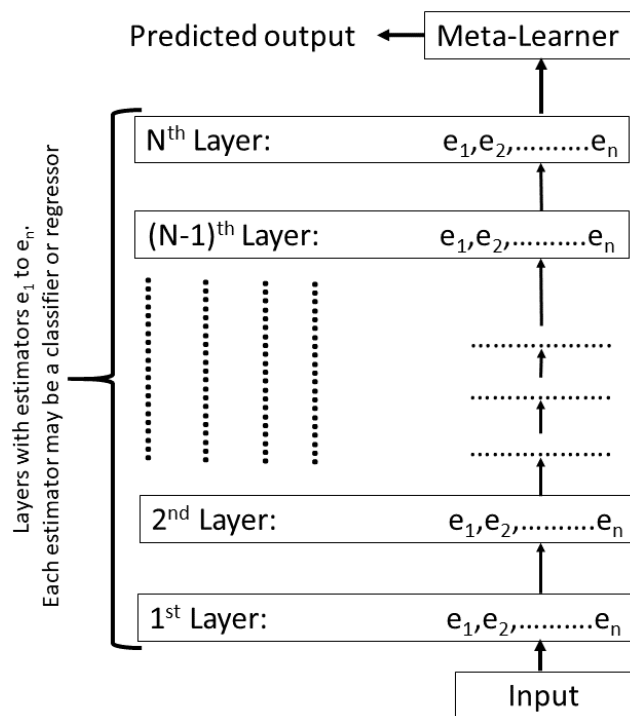


FIGURE 3. Architecture of stacked ensemble learning model.

layer is run in parallel whereas estimators between the layers are run in sequential. Finally, all the estimators across the layers are stacked and the output is fed to the meta-learner as shown in figure 3 for the generation of predicted output.

The proposed Multi Layer Stacked Ensemble Learning Model (MLSELM) consists of four phases such as 1. Input phase which takes Phishing dataset as input, 2. Data balancing phase which is as described in section III-B and finally, 3. implementation of MLSELM as described in III-C.

A. DATASET

The proposed work applied on four datasets named as D1, D2, D3, D4. D1 is collected from UCI repository [17] and D2 is collected from mendely [20] with 48 features. Finally, D3 and D4 are also collected from mendely [30] where the dataset D3 contains 111 features with 58,645 instances, D4 contains with 111 features with 88,647 instances. Each dataset consists two classes such as phishing and legitimate. The description of each dataset is also shown in table 1.

From the description it is observed that the UCI phishing dataset [17] and two variants of mendely [30] are imbalanced. In this connection, as described in section III-B the data re-sampling method is applied to make the datasets as balanced to improve performance of the proposed model.

B. DATA BALANCING

In general, Random Under Sampling (RUS) and Random Over Sampling (ROS) techniques are applied on an imbalanced datasets to make it balanced. The ROS balancing method makes equal number of instances for both minority and majority class. This method randomly duplicate the instances of minority class and adds to itself to make minority

TABLE 1. Description of both datasets.

Sno	Dataset	Description
1	D1 [18]	It consists 30 features with 11055 instances. In which, 4898 are legitimate and 6157 are phishing. The legitimate indicated with 1 and phishing with 0 .
2	D2 [21]	It has 48 features and 10,000 instances. In which, 5000 are legitimate and 5000 are phishing. The legitimate and phishing are labeled as 1 and 0 .
3	D3 [31]	It has 111 features and 58,645 instances. In which, 30,647 are legitimate and 27,998 are phishing. The legitimate and phishing are labeled as 0 and 1 .
4	D4 [31]	It has 111 features and 88,647 instances. In which, 30,647 are legitimate and 58,000 are phishing. The legitimate and phishing indicated with 0 and 1 .

class instances equal with majority class. We apply data balancing on D1, D3 and D4 as they contain imbalanced data.

The D1 dataset has 4898 legitimate instances and 6157 phishing instances in which, legitimate class is treated as minority category and phishing as majority category. Hence, instances of legitimate category randomly duplicated and added to itself with ROS method to make equal number of instances for both minority and majority class. After balancing total number of instances are 12314 in which, legitimate are 6157 and phishing are 6157.

The phishing class of D3 identified as minority class since, it consists of 27998 instances. The legitimate with 30647 instances as treated majority class. The phishing class instances randomly duplicated and added to itself to make 27998 instances as 30647 by ROS data balancing method. Similarly, in D4 the legitimate class with 30647 instances identified as minority class and phishing with 58000 as majority class. The legitimate class instances randomly duplicated and added to itself to make 30647 as 58000 instances. After data balancing the total number of instances in D3 are 61294 and 116000 instances in D4.

C. MLSELM

The working of the proposed model is shown in figure 2 where it includes multiple layers and last layer as meta learner. The classification algorithms such as XGB, LR, RF, MLP, and KNN are encompassed in first layer, The XGB, RF and MLP are combined in second layer and finally, XGB act as Meta-layer. The Layer-I, Layer-II, and meta layer are stacked. The main intention of choosing these classifiers is due to the existence of wide usage in literature and also the classifiers provide diverse working in classifying the data. The second layer classifiers are chosen such that best 3 classifiers from first layer are selected. First layer takes phishing dataset as input to all five classifier after applying balancing process on phishing dataset. Then, passes the result of the first layer to all classifiers of second layer and then, the results of layer-II are passed to meta layer. The results of this model on four phishing datasets are tabulated in table 2, 3, 4, and 5. The results section IV covers detailed discussion.

The learning algorithms encompassed in each layer-I, layer-II & layer-III are entirely different and unique when compared with each other as described below.

IV. EXPERIMENTATION RESULTS

The proposed MLSELM and Machine Learning algorithms such as MLP, KNN, RF, LR and XGB applied on four datasets shown in Table 1. The classification metrics such as Precision, Recall, F-score and Accuracy are considered to evaluate its performance. Here, in the context of Legitimate and Phishing, we term phishing instances as positive and legitimate instances as negative. The number of True Positive, True Negatives, False Positives and False Negatives are termed as follows:

- P: Total number of phishing instances
- N: Total number of legitimate instances
- N_{TN} : The number of legitimate instances predicted as legitimate
- N_{FN} : The number of phishing predicted as legitimate
- N_{TP} : The number of phishing instances predicted as phishing
- N_{FP} : The number of legitimate predicted as phishing

The calculation of each metric is as follows:

- Precision = $\frac{N_{TP}}{N_{TP}+N_{FP}} \times 100$
- Recall = $\frac{N_{TP}}{N_{TP}+N_{FN}} \times 100$
- F-score = $\frac{\text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \times 100$
- Accuracy = $\frac{N_{TP}+N_{TN}}{P+N} \times 100$

The results of MLSELM are compared with all five classification models with and without data balancing as described in section IV-A. Similarly, section IV-B analyses the results of MLSELM on four phishing datasets by comparing with the existing literature.

A. THE COMPARISON OF ML ALGORITHMS WITH MLSELM

In this section, we evaluate the performance of the proposed model and includes comparison of proposed work with various classifiers. The experimentation of this section is given below.

TABLE 2. The performance of various classifiers & MLSELM algorithm with and without data balancing on D1 dataset.

	Measures	XGB	LR	RF	MLP	KNN	MLSELM
Without Data Balancing	Accuracy	94.39	92.4	96.42	96.74	95.65	97.06
	Precision	91.42	90.37	94.56	94.87	94.45	95.08
	Recall	95.41	91.91	97.09	97.52	95.45	98.05
	F-Score	93.37	91.13	95.81	96.82	94.95	96.54
With Data Balancing ROS	Accuracy	97.19	92.12	97.56	97.07	96.58	97.76
	Precision	97.0	91.19	96.92	96.42	96.76	97.34
	Recall	97.25	92.58	98.06	97.56	96.52	98.07
	F-Score	97.13	91.88	97.49	96.99	96.28	97.70
With Data Balancing RUS	Accuracy	95.76	92.44	96.53	95.86	95.86	96.78
	Precision	95.16	91.41	96.15	95.16	95.75	96.54
	Recall	96.59	93.81	97.1	96.78	96.23	97.21
	F-Score	95.87	92.6	96.62	95.96	95.99	96.87

TABLE 3. The performance of various classifiers & MLSELM algorithm with and without data balancing on D2 dataset.

	Measures	XGB	LR	RF	MLP	KNN	MLSELM
Without Data Balancing	Accuracy	96.85	94.54	98.20	96.55	87.35	98.90
	Precision	96.65	92.91	98.07	94.43	83.50	98.48
	Recall	96.95	94.54	98.27	98.52	90.16	99.28
	F-Score	96.80	93.72	98.17	96.43	86.70	98.88

TABLE 4. The performance of various classifiers & MLSELM algorithm with and without data balancing on D3 dataset.

	Measures	XGB	LR	RF	MLP	KNN	MLSELM
Without Data Balancing	Accuracy	92.7	88.6	94.85	86.47	87.37	96.5
	Precision	90.02	86.29	93.83	90.86	85.36	96.42
	Recall	94.41	89.44	95.31	82.54	87.82	96.25
	F-Score	92.16	87.83	94.57	86.5	86.57	96.33
With Data Balancing ROS	Accuracy	95.99	87.69	95.02	84.28	88.22	96.79
	Precision	95.68	84.33	94.22	90.48	87.26	96.84
	Recall	96.23	90.28	95.70	80.37	88.84	96.70
	F-Score	95.96	87.20	94.95	85.12	88.04	96.77
With Data Balancing RUS	Accuracy	95.1	90.24	94.97	94.43	93.75	96.64
	Precision	94.68	89.12	94.3	95	94.14	96.73
	Recall	95.49	91.16	95.58	93.94	93.41	96.56
	F-Score	95.08	90.13	94.94	94.47	93.77	96.64

The proposed model is evaluated by applying on different datasets(D1,D2,D3,D4) with and without data balancing. The results with and without data balancing are shown with respect to different datasets. Table 2 shows the results of proposed model with dataset D1 with and without data balancing. From the results, it is observed that the proposed model achieved a significant accuracy of 97.76% with ROS data balancing outperforming existing classifiers and MLSELM without data balancing.

Similarly, proposed model is applied on D2 to detect phishing site. Table 3 gives the performance results of the proposed model. As the dataset is already balanced, we donot include

data balancing on D2. From the results, it is clearly seen that, MLSELM outperformed other classifiers with a significant accuracy of 98.90%, precision of 98.48% and F-score of 98.88%.

The results with D3 dataset is given in Table 4 which shows the performance of the model with and without data balancing. From the table, it is demonstrated that MLSELM with ROS achieved an accuracy of 96.79% outperforming MLSELM on imbalanced data and other classifiers.

Finally, the proposed model is applied on dataset D4 to result Table 5. The table gives the result with and without data balancing. From the results, it is observed that MLSELM

TABLE 5. The performance of various classifiers & MLSELM algorithm with and without data balancing on D4 dataset.

	Measures	XGB	LR	RF	MLP	KNN	MLSELM
Without Data Balancing	Accuracy	95.17	92.31	96.4	92.05	89.7	97.41
	Precision	94.86	92.71	96.75	92.67	92.24	97.88
	Recall	97.7	95.41	97.72	95.07	92.05	98.16
	F-Score	96.26	94.04	97.23	93.85	92.14	98.02
With Data Balancing ROS	Accuracy	97.35	92.28	96.78	90.10	92.63	98.43
	Precision	97.07	90.66	95.86	96.10	88.72	97.93
	Recall	97.67	93.84	97.72	85.95	96.40	98.96
	F-Score	97.37	92.22	96.78	90.74	92.40	98.44
With Data Balancing RUS	Accuracy	96.14	92.99	96.03	95.56	95.3	97.09
	Precision	95.28	91.41	94.77	94.78	94.74	96.24
	Recall	96.85	94.2	97.11	96.15	95.68	97.83
	F-Score	96.06	92.78	95.93	95.46	95.21	97.03

TABLE 6. Comparison of results with MLSELM and existing works EW1,EW2,EW3,EW4,EW5,EW6 on D1 Dataset.

Measures	EW1[36]	EW2[37]	EW3[40]	EW4[42]	EW5[43]	EW6[44]	MLSELM
Accuracy	97.16	97.33	93.41	95.0	97.4	95.4	97.76
Precision	96.86	97.0	92.99	95.0	96.0	93.5	97.34
Recall	96.83	98.3	91.98	95.0	98.1	95.9	98.07
F-Score	—	97.6	92.48	95.0	97.0	97.0	97.70

achieved a significant difference of 1% with respect to accuracy when compared with MLSELM with ROS without data balancing approach and other classifiers. From the results across all datasets, it is evident that the proposed model outperformed other classifiers with significant performance irrespective of inclusion or exclusion of data balancing techniques. Also, the proposed model with data balancing had achieved significant performance compared with MLSELM without data balancing.

B. THE COMPARISON OF MLSELM WITH EXISTING LITERATURE

In this section, we compare our proposed model with existing works. As we tested our model on various datasets, we compare our work with the existing works that used same datasets. For the dataset D1, we considered Al-Sarem *et al.* [35], Basit *et al.* [36], Nagaraj *et al.* [39], Taha [41], Zamir *et al.* [42], and Ubing *et al.* [43] as existed works and termed as EW1, EW2, EW3, EW4, EW5, and EW6 respectively. The comparison results are given in Table 6. From the results, it is clearly observed that MLSELM outperformed existing techniques with a significant accuracy of 97.76%, precision of 97.34% and F score of 97.70%. For the dataset D2, we considered [35] and [38] as existing works and termed as EW7 and EW8. The comparison with existing works is shown in Table 7. From the results, it is observed that MLSELM achieved a significant performance with accuracy of 98.90% outperforming existing works.

Similarly, for the dataset 3 and 4, we considered Al-Sarem *et al.* [35] as existing work and termed as EW9.

TABLE 7. Comparison of results with EW7,EW8 and MLSELM on D2 dataset.

Measures	EW7[36]	EW8[39]	MLSELM
Accuracy	98.57	97.30	98.90
Precision	98.50	—	98.48
Recall	98.64	—	99.28
F-Score	98.57	—	98.88

TABLE 8. Comparison of results with EW9 and MLSELM on D4 dataset.

Measures	EW9[36]	MLSELM
Accuracy	97.35	98.43
Precision	96.20	97.93
Recall	96.14	98.96
F-Score	96.17	98.44

Since there exists no work applied on D3 dataset, we could not include the comparison results with existing works. From the results shown in Table 8, it is observed that proposed model outperformed existing work with a significant difference and achieved an accuracy of 98.43%, precision of 97.93% and F Score of 98.44%.

V. DISCUSSION

The main intention of this work is to provide the phishing detection with a significant accuracy. To achieve the same, we have designed the model with better classifiers being selected at each layer to take the advantage of diverseness of the classifiers. We have experimented the model with

and without data balancing techniques for identifying the effectiveness of the model with the instances of minor and major classes. The data balancing techniques used in the proposed work include Random Under Sampling and Random Over Sampling. We have also considered various datasets with different sizes and feature space for the evaluation of MLSELM. The rationale behind choosing different sizes is to observe the behavior of MLSELM with varying data. It has been clearly observed that MLSELM performed better with data balancing technique compared with imbalanced data. For all the datasets except D3, it is demonstrated that proposed model achieved a Recall of atleast 98% indicating the effectiveness of detection of phishing sites. The precision of model with various datasets also ranges from 96% to 98%. MLSELM compared with various traditional algorithms such as XGB,LR,RF,MLP,KNN and observed that XGB and RF performed better than other traditional algorithms but performed lower than MLSELM. The Multi-layer stacked ensemble used MLP, KNN, RF, LR, XGB at layer 1, RF,XGB, MLP at layer 2 and XGB as meta learner. Note that, various combinations of classifiers at different layers are experimented to get the suitable and better classifiers for the corresponding layer and there by reaches the final result with significant performance. Eventhough, the proposed model MLSELM achieved significant performance in detecting phishing websites but it has certain limitations which are given below. One of the limitation of MLSELM is the increase of overhead due to the stacking of classifiers at multiple layers. However, the computation overhead can be reduced using parallelization through multiple cores in a system. Eventhough the classifiers with in the layer can be run in parallel but layers are supposed to be run in sequence.

VI. CONCLUSION

In this paper, we proposed a multi layer stacked ensemble model for the detection of phishing sites. Diverse classifiers are attempted at different layers to achieve better performance compared to weak learners. The proposed model with D1,D2,D3,D4 datasets achieved an accuracy of 97.76%, 98.90%, 96.79% and 98.43% respectively. From the results, it is observed that MLSELM worked better with balanced data compared to imbalanced data. Moreover, the proposed model outperformed various baseline models and achieved significant difference in various evaluation metrics. Also, the proposed model with various datasets with different quantities are attempted and could achieve atleast 96.5% across all datasets. In future, we would like to use various feature selection algorithms to identify the significant features from the given datasets. Also, the fusion of feature selection algorithms with tuning parameters will be explored to improve the performance of the model.

REFERENCES

[1] M. Somesha, A. R. Pais, R. S. Rao, and V. S. Rathour, "Efficient deep learning techniques for the detection of phishing websites," *Sādhanā*, vol. 45, no. 1, pp. 1–18, Dec. 2020.

[2] R. S. Rao and A. R. Pais, "Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 9, pp. 3853–3872, Sep. 2020.

[3] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in *Proc. NDSS*, 2010, pp. 1–14. [Online]. Available: <http://www.isoc.org/isoc/conferences/ndss/10/pdf/08.pdf>

[4] R. S. Rao and A. R. Pais, "An enhanced blacklist method to detect phishing websites," in *Proc. Int. Conf. Inf. Syst. Secur.* Mumbai, India: Springer, 2017, pp. 323–333.

[5] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 1245–1254.

[6] R. Srinivasa Rao and A. R. Pais, "Detecting phishing websites using automation of human behavior," in *Proc. 3rd ACM Workshop Cyber-Physical Syst. Secur.* New York, NY, USA: ACM, Apr. 2017, pp. 33–42, doi: 10.1145/3055186.3055188.

[7] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell, "Client-side defense against web-based identity theft," in *Proc. NDSS*, 2004, pp. 1–16. [Online]. Available: <http://www.isoc.org/isoc/conferences/ndss/04/proceedings/Papers/Chou.pdf>

[8] H. Shahriar and M. Zulkernine, "Trustworthiness testing of phishing websites: A behavior model-based approach," *Future Gener. Comput. Syst.*, vol. 28, no. 8, pp. 1258–1271, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167739X11000045>

[9] R. S. Rao and A. R. Pais, "Jail-phish: An improved search engine based phishing detection system," *Comput. Secur.*, vol. 83, pp. 246–267, Jun. 2019.

[10] A. V. Ramana, K. L. Rao, and R. S. Rao, "Stop-phish: An intelligent phishing detection method using feature selection ensemble," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–9, Dec. 2021.

[11] R. S. Rao, T. Vaishnavi, and A. R. Pais, "CatchPhish: Detection of phishing websites by inspecting URLs," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 2, pp. 813–825, Feb. 2020.

[12] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing web sites," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 2, p. 21, 2011. [Online]. Available: <http://dl.acm.org/citation.cfm?doi=2019599.2019606>

[13] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Comput. Appl.*, vol. 31, no. 8, pp. 3851–3873, Aug. 2019.

[14] M. Khoraji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2091–2121, 4th Quart, 2013.

[15] V. Shahrivari, M. Mahdi Darabi, and M. Izadi, "Phishing detection using machine learning techniques," 2020, *arXiv:2009.11116*.

[16] S. Priya, S. Selvakumar, and R. L. Velusamy, "Gravitational search based feature selection for enhanced phishing websites detection," in *Proc. 2nd Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Mar. 2020, pp. 453–458.

[17] R. Mohammad, L. McCluskey, and F. Thabtah. *UCI Machine Learning Repository: Phishing Websites Data Set (2015)*. Accessed: 2016. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Phishing+Websites>

[18] S. Adi, Y. Pristyanto, and A. Sunyoto, "The best features selection method and relevance variable for web phishing classification," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIAC)*, Jul. 2019, pp. 578–583.

[19] M. Almseidin, A. Abu Zuraiq, M. Al-kassabeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," *Int. J. Interact. Mobile Technol.*, vol. 13, no. 12, p. 171, Dec. 2019.

[20] C. L. Tan, "Phishing dataset for machine learning: Feature evaluation," *Mendeley Data*, vol. 1, 2018. [Online]. Available: <https://data.mendeley.com/datasets/h3cgnj8hft/1>

[21] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," *Soft Comput.*, vol. 23, no. 12, pp. 4315–4327, Jun. 2019.

[22] S.-M. Javadi-Moghaddam and M. Golami, "Detecting phishing pages using the relief feature selection and multiple classifiers," *Int. J. Electron. Secur. Digit. Forensics*, vol. 12, no. 2, pp. 229–242, 2020.

[23] B. Alotaibi and M. Alotaibi, "Consensus and majority vote feature selection methods and a detection technique for web phishing," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 1, pp. 717–727, Jan. 2021.

[24] E. Zhu, C. Ye, D. Liu, F. Liu, F. Wang, and X. Li, "An effective neural network phishing detection model based on optimal feature selection," in *Proc. IEEE Intl Conf Parallel Distrib. Process. Appl., Ubiquitous Comput. Commun., Big Data Cloud Comput., Social Comput. Netw., Sustain. Comput. Commun. (ISPA/IUCC/BDCLOUD/SocialCom/SustainCom)*, Dec. 2018, pp. 781–787.

- [25] E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu, "OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network," *IEEE Access*, vol. 7, pp. 73271–73284, 2019.
- [26] S. R. Sharma, R. Parthasarathy, and P. B. Honnavalli, "A feature selection comparative study for web phishing datasets," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECT)*, Jul. 2020, pp. 1–6.
- [27] A. Singh and A. Tiwari, "A study of feature selection and dimensionality reduction methods for classification-based phishing detection system," *Int. J. Inf. Retr. Res.*, vol. 11, no. 1, pp. 1–35, Jan. 2021.
- [28] E. Gandotra and D. Gupta, "An efficient approach for phishing detection using machine learning," in *Multimedia Security*. Cham, Switzerland: Springer, 2021, pp. 239–253.
- [29] X. Dong, X. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020.
- [30] G. Vrbancic, "Phishing websites dataset," *Mendeley Data*, vol. 1, 2020. [Online]. Available: <https://data.mendeley.com/datasets/72ptz43s9v/1>
- [31] M. Al-Sarem, F. Saeed, A. Alsaedi, W. Boulila, and T. Al-Hadhrani, "Ensemble methods for instance-based Arabic language authorship attribution," *IEEE Access*, vol. 8, pp. 17331–17345, 2020.
- [32] X. Niu, J. Ma, Y. Wang, J. Zhang, H. Chen, and H. Tang, "A novel decomposition-ensemble learning model based on ensemble empirical mode decomposition and recurrent neural network for landslide displacement prediction," *Appl. Sci.*, vol. 11, no. 10, p. 4684, May 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/10/4684>
- [33] K. Yan, Y. Dai, M. Xu, and Y. Mo, "Tunnel surface settlement forecasting with ensemble learning," *Sustainability*, vol. 12, no. 1, p. 232, Dec. 2019. [Online]. Available: <https://www.mdpi.com/2071-1050/12/1/232>
- [34] P. Pintelas and I. E. Livieris, "Special issue on ensemble learning and applications," *Algorithms*, vol. 13, no. 6, p. 140, Jun. 2020. [Online]. Available: <https://www.mdpi.com/1999-4893/13/6/140>
- [35] M. Al-Sarem, F. Saeed, Z. G. Al-Mekhlafi, B. A. Mohammed, T. Al-Hadhrani, M. T. Alshammari, A. Alreshidi, and T. S. Alshammari, "An optimized stacking ensemble model for phishing websites detection," *Electronics*, vol. 10, no. 11, p. 1285, May 2021.
- [36] A. Basit, M. Zafar, A. R. Javed, and Z. Jalil, "A novel ensemble machine learning method to detect phishing attack," in *Proc. IEEE 23rd Int. Multi-topic Conf. (INMIC)*, Nov. 2020, pp. 1–5.
- [37] B. A. Tama and K.-H. Rhee, "A comparative study of phishing websites classification based on classifier ensemble," *J. Korea Multimedia Soc.*, vol. 21, no. 5, pp. 617–625, 2018.
- [38] P. Vaitkevicius and V. Marcinkevicius, "Composition of ensembles of recurrent neural networks for phishing websites detection," in *Proc. Int. Baltic Conf. Databases Inf. Syst.* Tallinn, Estonia: Springer, 2020, pp. 297–310.
- [39] K. Nagaraj, B. Bhattacharjee, A. Sridhar, and S. Gs, "Detection of phishing websites using a novel twofold ensemble model," *J. Syst. Inf. Technol.*, vol. 20, no. 3, pp. 321–357, Nov. 2018.
- [40] A. Niranjani, D. HariPriya, R. Pooja, S. Sarah, P. Deepa Shenoy, and K. Venugopal, "EKRV: Ensemble of KNN and random committee using voting for efficient classification of phishing," in *Progress in Advanced Computing and Intelligent Engineering*. Cham, Switzerland: Springer, 2019, pp. 403–414.
- [41] A. Taha, "Intelligent ensemble learning approach for phishing website detection based on weighted soft voting," *Mathematics*, vol. 9, no. 21, p. 2799, Nov. 2021.
- [42] A. Zamir, H. U. Khan, T. Iqbal, N. Yousaf, F. Aslam, A. Anjum, and M. Hamdani, "Phishing web site detection using diverse machine learning algorithms," *Electron. Library*, vol. 38, no. 1, pp. 65–80, Mar. 2020.
- [43] A. A. Ubung, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam, "Phishing website detection: An improved accuracy through feature selection and ensemble learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, pp. 252–257, 2019.
- [44] V. E. Adeyemo, A. O. Balogun, H. A. Mojeed, N. O. Akande, and K. S. Adewole, "Ensemble-based logistic model trees for website phishing detection," in *Proc. Int. Conf. Adv. Cyber Secur.* Penang, Malaysia: Springer, 2020, pp. 627–641.
- [45] A. Subasi and E. Kremic, "Comparison of AdaBoost with multiboosting for phishing website detection," *Proc. Comput. Sci.*, vol. 168, pp. 272–278, Jan. 2020.
- [46] Z. G. Al-Mekhlafi, B. A. Mohammed, M. Al-Sarem, F. Saeed, T. Al-Hadhrani, M. T. Alshammari, A. Alreshidi, and T. S. Alshammari, "Phishing websites detection by using optimized stacking ensemble model," *Comput. Syst. Sci. Eng.*, vol. 41, no. 1, pp. 109–125, 2022.



LAKSHMANA RAO KALABARIGE received the Ph.D. degree in the area of wireless and cognitive radio networks from Gitam University, Visakhapatnam. He is currently working as an Associate Professor at the GMR Institute of Technology, Rajam, India. His research interests include machine learning, deep learning, natural language processing, and computer vision.



ROUTHU SRINIVASA RAO received the B.Tech. degree in computer science and engineering from the SRKR Engineering College, Andhra University, India, the M.Tech. degree in computer science and engineering from NIT Kurukshetra, Haryana, India, and the Ph.D. degree in the area of cyber security from NITK Surathkal. He is currently working as an Associate Professor at the GITAM School of Technology, GITAM (Deemed to be University), Visakhapatnam, India. His research interests include information security, cyber security, phishing, machine learning, and natural language processing.



AJITH ABRAHAM received the Ph.D. degree in computer science from Monash University, Melbourne, Australia, in 2001, and the Master of Science degree from Nanyang Technological University, Singapore, in 1998. He is currently the Director of the Machine Intelligence Research Laboratories (MIR Laboratories), a Not-for-Profit Scientific Network for Innovation and Research Excellence connecting Industry and Academia. The Network with HQ in Seattle, USA, has currently more than 1,500 scientific members from over 105 countries. As an Investigator/Co-Investigator, he has won research grants worth over more than 100 Million USD. He is also working as a Professor in artificial intelligence with Innopolis University, Russia, and is the Chair Holder of the Yayasan Tun Ismail Mohamed Ali Professorial Chair in Artificial Intelligence of UCSI, Malaysia. He works in a multi-disciplinary environment and he has authored/coauthored more than 1,400 research publications out of which there are more than 100 books covering various aspects of computer science. One of his books was translated to Japanese and few other articles were translated to Russian and Chinese. He has more than 46,000 academic citations (H-index of more than 102 as per google scholar). He has given more than 150 plenary lectures and conference tutorials (in more than 20 countries). Since 2008, he has been the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over more than 200 members), during 2008 to 2021, and served as a Distinguished Lecturer of IEEE Computer Society representing Europe (2011–2013). He was the Editor-in-Chief of *Engineering Applications of Artificial Intelligence* (EAAI), during 2016 to 2021, and serves/served the editorial board of over 15 international journals indexed by Thomson ISI.



LUBNA ABDELKAREIM GABRALLA received the B.Sc. and M.Sc. degrees in computer science from the University of Khartoum and the Ph.D. degree in computer science from the Sudan University of Science and Technology, Khartoum, Sudan. She became a Senior Fellow (SFHEA), in 2021. She is currently an Associate Professor with the Department of Computer Science and Information Technology, Princess Nourah Bint Abdulrahman University, Saudi Arabia. Her current research interests include soft computing, machine learning, and deep learning.

...