IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 June 2023, accepted 6 July 2023, date of publication 10 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293649

RESEARCH ARTICLE

A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites

LAKSHMANA RAO KALABARIGE[1], ROUTHU SRINIVASA RAO[2], ALWYN R. PAIS[3], AND LUBNA ABDELKAREIM GABRALLA[4]

[1]AI Research Laboratory, GMR Institute of Technology, Rajam 532127, India
[2]Department of Computer Science and Engineering, Gandhi Institute of Technology and Management, Visakhapatnam, Andhra Pradesh 530045, India
[3]Information Security Research Laboratory, Department of Computer Science and Engineering, National Institute of Technology, Surathkal, Karnataka 575025, India
[4]Department 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 research was funded by Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R178), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT Phishing is a type of online scam where the attacker tries to trick you into giving away your personal information, such as passwords or credit card details, by posing as a trustworthy entity like a bank, email provider, or social media site. These attacks have been around for a long time and unfortunately, they continue to be a common threat. In this paper, we propose a boosting based multi layer stacked ensemble learning model that uses hybrid feature selection technique to select the relevant features for the classification. The dataset with selected features are sent to various classifiers at different layers where the predictions of lower layers are fed as input to the upper layers for the phishing detection. From the experimental analysis, it is observed that the proposed model achieved an accuracy ranging from 96.16 to 98.95% without feature selection across different datasets and also achieved an accuracy ranging from 96.18 to 98.80% with feature selection. The proposed model is compared with baseline models and it has outperformed the existing models with a significant difference.

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

I. INTRODUCTION

In recent times, the internet has brought about revolutionary changes in the way we communicate, making it more convenient and accessible. However, this positive transformation has also led to a significant increase in the number of internet users, providing an opportunity for adversaries to exploit naive individuals by stealing their sensitive credentials. One of the most common methods used by these attackers is phishing, which involves sending fake emails or creating replica websites to lure unsuspecting users into providing their personal information. As a result, innocent users become prey to these attacks.

Based on a phishing survey conducted by the Anti-Phishing Working Group (APWG),[1] the total number of phishing websites for the first three quarters of 2022 was 3,394,662, as depicted in Figure 1. In contrast, the combined total for all four quarters of 2021 was 2,847,773, as illustrated in Figure 2. This represents a significant 19.2% increase in just the first three quarters of 2022 when compared to the entire year of 2021. This growth highlights the serious threat posed to naive internet users. Phishing attacks are commonly developed and distributed over the internet through two methods: fake emails and the replication of legitimate websites. Fake emails, also known as spoofed emails, are sent to users under the guise of a legitimate company or organization. In addition,

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

[1]https://apwg.org/trendsreports/

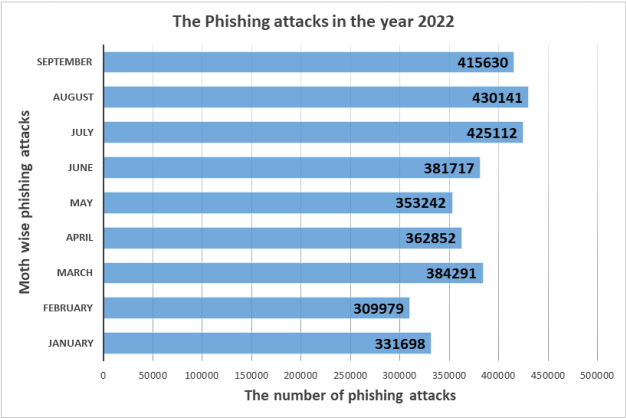

FIGURE 1. The trend of phishing attacks of first three quarters of the year 2022.

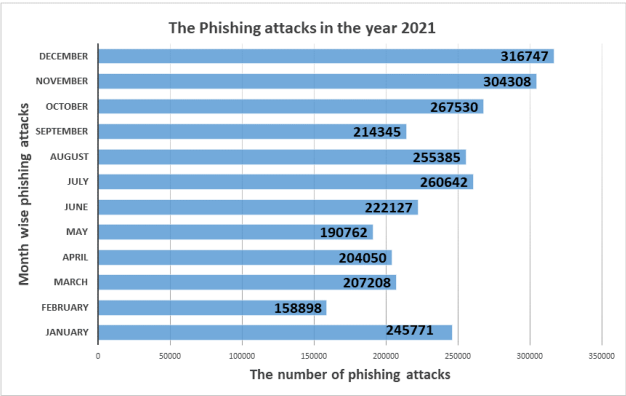

FIGURE 2. The trend of phishing attacks of all four quarters of the year 2021.

attackers create and deploy replicas of original websites on social media platforms like Twitter, Facebook, and Google. These phishing websites may use the green padlock and Hypertext Transfer Protocol Secure (HTTPS) to trick users into believing they are legitimate sites.

To detect and prevent phishing attacks, various methods have been proposed in the literature, including blacklist [1], [2], [3], [4], feature extraction [5], [6], [7], [8], and machine learning [9], [10], [11]. A blacklist is a list or database of phishing URLs that are typically blocked by modern browsers such as Chrome, Opera, and Mozilla. However, this technique is ineffective in detecting and preventing zero-day phishing sites that have a short lifespan. Feature extraction involves extracting characteristics from different phishing websites to identify and prevent phishing attacks, but not all phishing sites have the same features, so this method may not be reliable for all websites.

As a result, classification models [12] like Decision Tree (DT), Random forest (RF), etc. are used to detect phishing attacks. Existing literature [2], [12], [13] shows that machine learning-based methods can achieve up to 99% accuracy in detecting phishing websites, outperforming the blacklist and feature extraction techniques.

The performance of machine learning (ML) algorithms for detecting and preventing phishing attacks depends on the quantity of training data and the quality of the extracted features from phishing websites. Traditional ML models struggle to capture the diverse characteristics of data, while ensemble learning can extract diversified features, combine predictive results produced by multiple learning algorithms, and achieve better predictive performance using ensemble methods like voting, stacking, blending, and averaging. Also, deep learning methods are used in different domains [38], [39], [40], [41] including medical, security and NLP. Techniques in [41], [42], [43], and [44] use different deep learning techniques such as LSTM, CNN, GRU etc for the classification of phishing sites. To further select the relevant features from the given dataset, feature selection algorithms such as filter, wrapper and embedded techniques are used.

This work proposes a feature selection-based ensemble model to detect and prevent phishing websites, aiming to reduce the time for training and classification, as well as computation overhead. By harnessing the capabilities of a range of well-performed models in the task of classification, the proposed ensemble model shows promise for detecting and preventing phishing attacks. The model is applied to four datasets, including two variants of the Mendeley Phishing Dataset (MPD) (small and large), Mendeley with 10,000 instances, and UCI.

A. MOTIVATION

Phishing attacks pose a significant threat to online security and detecting them accurately remains a challenging problem. Various machine learning and feature selection strategies have been proposed to address this issue. Baseline machine learning approaches have successfully identified phishing websites, but ensemble-based models have demonstrated better efficiency and accuracy.

Specifically, the stacking model MLSELM [45] achieved the best results among the baseline and ensemble models. Feature selection approaches have also been employed to obtain an optimal feature subset, reducing model execution time, and improving accuracy. Feature importance-based approaches have shown greater accuracy, as they rank each feature based on its contribution to the model. However, these approaches have not been fully explored, especially in boosting-based ensemble models, stacking, multi-layered stacking, and the averaging of feature ranks obtained from multiple boosting models.

B. CONTRIBUTION

This study proposes a novel hybrid feature selection approach and a boosting-based multi-layered stacking ensemble learning model to address the challenges of detecting phishing attacks accurately. The feature importance ranking of three out of five boosting models that achieved high accuracy on all

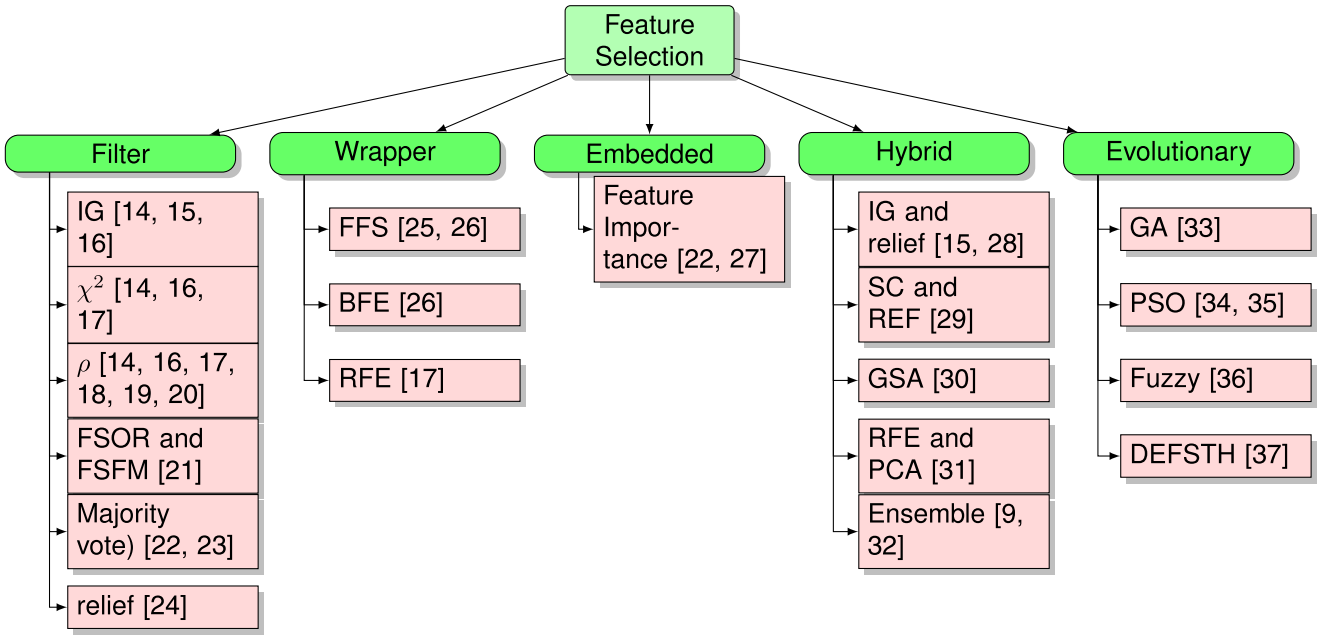

FIGURE 3. Different feature selection approaches.

four phishing datasets were considered. The average feature subset was determined from the three feature subsets selected by the three best boosting models for each dataset. Finally, K-topmost features were selected for each dataset, with K ranging from 66% to 86% based on the number of features and size of the respective dataset.

The proposed multi-layered stacking model integrates the four best-performing boosting models as in the architecture of previously developed MLSELM [45] model. The model achieved high accuracy on all four phishing datasets using all features, both for imbalanced and balanced data.

Additionally, the hybrid feature selection approach followed by boosting-based multi-layered stacking model achieved high accuracy for both imbalanced and balanced data with reduced features. The hybrid feature selection approach identified the most informative features for detecting phishing attacks accurately, reducing the number of features used in the models. The results demonstrate that the proposed approach achieves high accuracy while using a reduced number of features. The proposed model is designed to achieve significant detection rate using hybrid feature selection and Multi-layered stacked ensemble model. Boosting focuses on reducing bias, while the stacking framework combines the strengths of different models. This combination helps mitigate the weaknesses of individual models, leading to improved generalization performance on unseen data. It is evaluated on different datasets to evaluate the behavior of the model with varying datasets. The model can be deployed as a web application or a browser extension which takes input as URL and source code of the websites and can result the web page as either legitimate or phishing.

The organization of the remaining of this paper is as follows. In section II, a review of literature on feature selection-based phishing detection and prevention techniques is presented. Section III outlines the architecture and functionality of the proposed model and covers the implementation of various phases of the proposed model, including the input dataset and the feature selection ensemble model. Section IV presents the experimental results with both the baseline and ensemble models as well as provides justifications, key findings, and limitations of the proposed model. Lastly, section V concludes the paper.

II. RELATED WORK

The feature selection techniques were classified into different categories as shown in Figure 3. such as 1. Filter, 2. Wrapper, 3. Embedded, 4. Hybrid, and 5. Evolutionary. The Information Gain (IG), Chi-square test(χ^2), Fisher's score, Correlation Coefficient(ρ), Variance Threshold, mean absolute difference (MAD), relief (reliefF, RreliefF) and Dispersion Ratio are knowing as Filter methods. Whereas, Forward Feature Selection (FFS), Backward Feature Elimination (BFE), Exhaustive Feature Selection (EFS), and Recursive Feature Elimination (RFE) are known as Wrapper methods. On the other hand, LASSO Regularization(L1), and Feature Importance are known as Embedded approaches. The combination of more than one feature selection approach is known as hybrid and evolutionary-based feature selection is a category of wrapper approach used to select optimal feature subset through evolutionary algorithms.

The below are some the proposed feature selection approaches belongs to either Filter, Wrapper or Embedded. These three types of approaches were applied in different

TABLE 1. Description of datasets and selected features through hybrid feature selection.

Sno	Dataset	Description	Selected K-topmost features
1	**D1** [46]	The dataset consists of **11,055** instances with 30 features. Of these instances, **4,898** are labeled as legitimate (indicated by **1**), and **6,157** are labeled as phishing (indicated by **0**).	**22** features were selected, which is equal to **66**% of total features.
2	**D2** [47]	The dataset comprises **10,000** instances with **48** features. Half of these instances are labeled as legitimate (indicated by **1**), while the other half are labeled as phishing (indicated by **0**).	**33** features were selected, which is almost **68**% of total features
3	**D3** [48]	The dataset contains **58,645** instances for **111** features. Of these instances, **30,647** are labeled as phishing (indicated by **1**), and **27,998** are labeled as legitimate (indicated by **0**).	**96** features were selected, which is about **86**% of total features
4	**D4** [48]	The dataset contains **88,647** instances for **111** features. Of these instances, **30,647** are labeled as phishing (indicated by **1**), and **58,000** are labeled as legitimate (indicated by **0**).	**83** features were selected, which is near to**75**% of total features

combinations on four (D1, D2, D3, and D4) phishing datasets as described below. The filter method ReliefF [24], applied on UCI and selected 17 features. The ReliefF followed by Majority voting on multiple baseline classification approach obtained 95% accuracy. Furthermore, correlation feature selection (CFS) [18] selected 23 features from the UCI phishing dataset and CFS followed by statistical t-test with KNN obtained 97% of accuracy. Where, highly correlated features were considered as redundant and removed by CFS and the significance of features were tested through statistical t-test to obtain the most relevant features. Likewise, [14], applied four filter-based feature selection methods such as Correlation-Based Features Selection (CBFS), Information Gain (IG), Information Gain Ratio (Gain Ratio), and Chi-Square on UCI phishing dataset. Each FS approach selected 9 different features respectively. It is observed that the accuracy of baseline models, namely Naive Bayes (NB), Decision Tree (ID3 And C4.5), K-Nearest Neighbour (KNN), and Support Vector Machine is decreased and obtained accuracy within range of 94.01% to 94.17%. Moreover, in [15], selected 20 features from Mendeley [47] through IG and ReliefF approaches. The FS approaches followed by RF obtained 98.11% accuracy. In a similar way, Union of IG and relief using RF [28] with 20 features obtained 98.11% accuracy. In addition, Prince et al. [16] compared and analysed multiple feature selection methods: Chi-Square, Gain Ratio(GR), Information Gain(IG), Pearson Correlation Coefficient(PCC), and Principal Components Analysis(PCA). Among all FS methods, Info Gain with Random Forest on 32 feature subset acquires 98.38% accuracy. Likewise, [17], employed four FS approaches, namely Chi-Squared, Keiser-Meyer-Olkin (KMO), Recursive Feature Elimination(RFE), and Pearson Correlation used along with RF, DT, SVM, KNN, and Multi-Layer perceptron(MLP) baseline models on both UCI and Mendeley [47] phishing datasets. The Chi-Squared followed by RF with 15 features and Pearson Correlation followed by RF with 14 features obtained 96.2% accuracy on UCI dataset. On the other hand, REF followed by RF with 26 features obtains 97.8% accuracy on Mendeley dataset. Likewise, Karabatak and Mustafa [26], applied five wrapper based FS techniques such as Individual Feature Selection (IFS), Forward Feature Selection (FFS), Backward Feature Selection (BFS), Plus-l takeaway-r FS (l=3, r=1), and Association Rule(AR) on UCI. The selected features were 27,24,25,27, and 26 respectively. In which, AR with RF obtained 97.31% accuracy. Furthermore, Abdulrahaman et al. [25] employed Wrapper Subset Evaluator with Ranker (WSER) method on UCI dataset selected 28 features and obtains 97.2953% accuracy through WSER followed by RF. The embedded approach such as Random Forest Regression (RFR) [22] based feature importance approach selected 9 features among 30 features of UCI dataset and majority voting-based ensemble model obtained 95.4% accuracy with the selected nine feature subset.

Likewise, RF based feature importance [27] applied on Mendeley-full [48] dataset and selected 14 features out of 111 features and obtained 97% accuracy with RF.

In addition to filter, embedded and wrapper approaches the hybrid FS approach also employed by some researchers to obtain optimal features. The combination of more than one FS approach of same category or different category is known as Hybrid FS approach. Zamir et al. [31] applied combination of filter (information gain, gain ratio, Relief-F) and wrapper (recursive feature elimination (RFE)) based FS approaches on UCI which obtained 27 features and the normalized 27 feature subset fed to the Principal Component Analysis (PCA) followed by Stacking (NN+RF+Bagging) obtained 97.4% accuracy.

Likewise, Moedjahedy et al. [29] applied three combinations of hybrid FS approaches. The first combination was predictive score correlation (PSC) and REF, next, maximal information coefficient correlation (MICC) and REF, and finally, spearman correlation (SC) and REF. From the results, it was observed that the third combination SC and REF using RF with 10 features obtained 97.6% of accuracy.

In addition, the hybrid feature ensemble [9] employed five FS approaches, namely Info Gain(IG), ANOVA, RFE, ReliefF, and Fisher Score. The best performed three approaches IG, ANOVA and RFE are ensembled and achieved 97.51% of accuracy on UCI and 98.45% accuracy on Mendeley [47].

In recent times, evolutionary learning approaches gains attention by the researchers as another alternative approach to determine best feature subset. As a result, some of

the evolutionary algorithms applied on UCI, Mendeley, Mendeley-small, and Mendeley-full phishing datasets. In which some of them are as follows. The Gravitational Search Algorithm(GSA) [30] with Random Forest(RF) model obtained 95.53% accuracy. The GSA selected 15 features from UCI dataset and found that its performance is better than other feature section methods, namely Correlation Feature Selection (CFS), Information Gain (IG), and Principal Component analysis (PCA). Likewise, the wrapper method with Genetic Algorithm [33] using DT classifier applied on UCI dataset and selected 20 best features. The performance of selected features evaluated through Nonlinear Regression based Harmony Search (NR-HS) (meta-heuristic nonlinear regression approach) and SVM. The accuracy of these two models were 92.8% and 91.83% respectively. Moreover, Laplacian Particle Swarm Optimization (LAPPSO) [34] and Filter based Bare-bone Particle Swarm Optimization(FBPSO) applied on UCI and Mendeley [47] phishing datasets. The LAPPSO selects 20 features from UCI and 17 from Mendeley. On the other hand, FBSO selects 18 and 26 features respectively. The performance of FS LAPPSO and FBPSO compared through baseline models DT and KNN. The DT with LAPPSO secures 96.6% on UCI and 95.8% of F-score on UCI and Mendeley [47]. datasets respectively.

In addition, fuzzy rough set (FRS) [36] selected 24 and 30 features respectively from UCI and Mendeley [47] phishing datasets. The FRS followed by RF obtained 93% and 95% of F-score. Moreover, differential evolution for feature selection with threshold mechanism (DEFSTH) [37] followed by Naïve Bayes classifier applied on on Mendeley-Full dataset and obtained 96.82% of accuracy.

Likewise, Binary Slap Swarm Optimization Algorithm (BSSA) [35] with transfer functions(TF) such as S-shaped, U-shaped, V-shaped, X-shaped, and Z-shaped TFs were applied on Mendeley(111 features) phishing dataset and selected 49 best feature subset among 111 features. From the results it is observed that the BSSA with X-shaped TF followed by KNN outperforms all other TFs with 95.07% accuracy. Similarly, some approaches other than filter, wrapper, embedded, hybrid, and evolutionary approaches applied to obtain optimal feature subset from phishing datasets. The Hybrid Ensemble Feature Selection (HEFS) [32], applied hybrid perturbation ensemble (i.e., data perturbation and function perturbation) followed by Cumulative Distribution Function gradient (CDF-g) for automatic feature cut-off rank identification approach to obtain final feature subset. The HEFS selected 10 features and HEFS followed by RF obtains 94.6% of accuracy. Likewise, Effective Neural Network Phishing Detection Model Based on Optimal Feature Selection (OFS-NN) [19] approach applied on UCI and obtained 96.75% of accuracy with 26 feature subset. Likewise, [20], applied feature validity value (FVV) index select the optimal features from UCI phishing dataset. The FVV obtained 23 features and FVV followed by NN obtained 94.5% of accuracy. Moreover, two FS approaches, namely Feature Selection by Omitting Redundant Features(FSOR) and Feature Selection by Filtering Method (FSFM) [21] applied on UCI dataset. The FSOR followed by RF with 22 features obtained 97.18% accuracy and FSFM followed by RF with 9 features obtained 95.21% accuracy. Furthermore, the eighteen common features of UCI and Mendeley [47] datasets were combined in [49] and selected 13 optimal features among 18 through Variance inflation factor (VIF) and P-Value feature analysis approaches. The RF on those 13 features achieved 93.2% of accuracy. Likewise, two feature selection approaches, namely consensus and majority voting [23] on UCI and Mendeley [47] phishing datasets. From Mendeley, 17 features were selected and obtained 98.17% of accuracy by consensus FS approach; 23 features were selected and obtained 98.63% of accuracy by majority voting FS. Likewise, from UCI, 9 features were selected and obtained 93.55% of accuracy by consensus approach and 13 features were selected and obtained 95.29% of accuracy by majority voting approach.

Majority of the existing works either used classical machine learning algorithms or ensemble algorithms (bagging and boosting) for the classification of algorithms. Some of the techniques also used feature selection algorithms such as filters or wrappers for identifying the relevant and significant features for the classification task. The proposed work uses different boosting algorithms for identifying the significant features using embedded method. The model also consists of multi layered stacked ensemble where stacked ensemble increases the model diversity and multi-layered structure enables hierarchical feature learning which learn different levels of abstractions from the data.

III. PROPOSED MODEL

The proposed work introduces a Boosting based Multi-layer stacked ensemble learning model (BMLSELM) to detect phishing websites. The model BMLSELM built based on MLSELM [45] using all boosting algorithms and it also uses hybrid feature selection method to select an optimal feature subset. It has three layers, as shown in figure 4, with all boosting algorithms. The first layer includes four estimators, namely XGBoost, LGBM, CatB, and AdaB. The second layer has three estimators, XGB, CatB, and AdaB, while XGB serves as the meta-learner in the final layer. Additionally, the hybrid feature selection method extracts essential features using three boosting models (XGB, CatB, and LGBM), finds feature ranking for all features through XGB, CatB, and LGBM, takes respective feature wise average for all three selected feature subset based on their feature ranks, and finally selects the K-topmost feature subset, which provides the highest accuracy, as presented in figure 5. The table 1 shows optimal percentage and number of K-topmost features from each phishing dataset.

The proposed approach involves four phases for evaluating phishing datasets. In the first phase, four phishing datasets were evaluated using five boosting and BMLSELM models. The second phase involved selecting the K-topmost features.

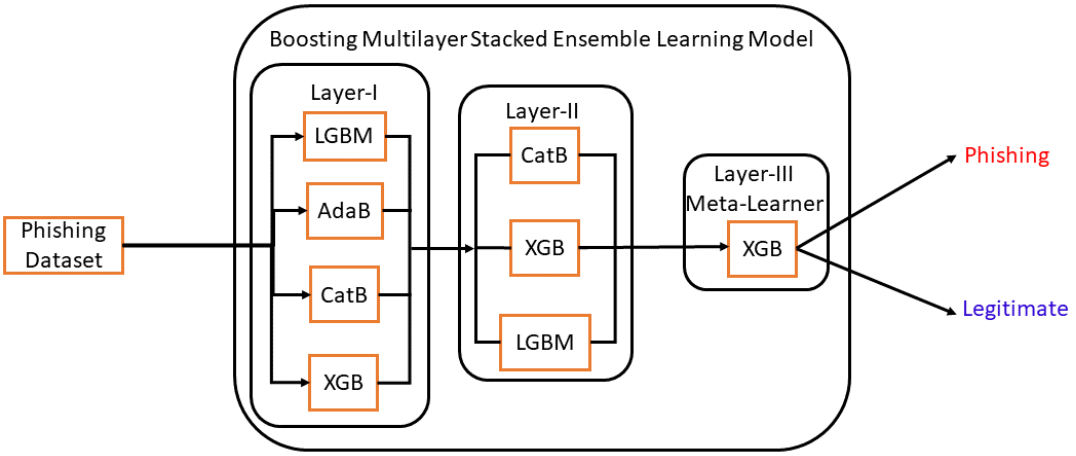

FIGURE 4. Architecture of boosting based multi-layer stacked ensemble learning model.

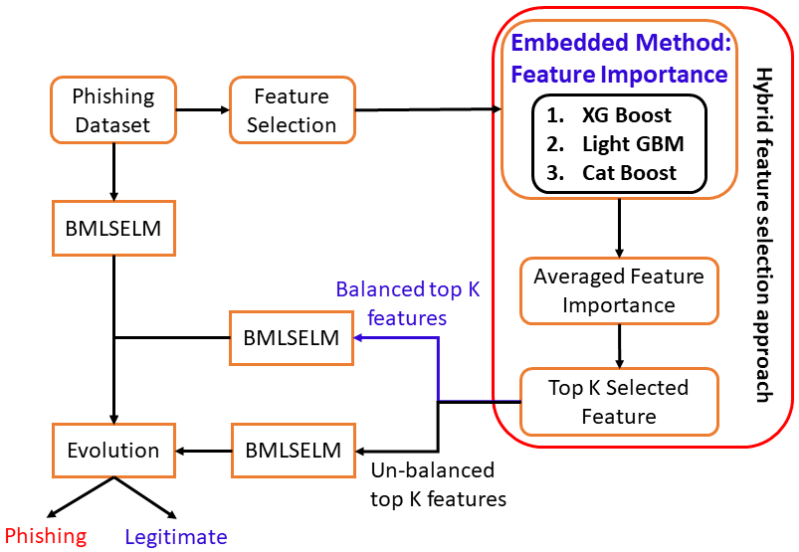

FIGURE 5. Different phases in boosting based multi-layer stacked ensemble learning model.

In the third phase, the unbalanced K-topmost features were evaluated using five boosting and BMLSELM models. Finally, the balanced K-topmost features were evaluated using five boosting and BMLSELM models as shown in figure 5.

A. DATASET

The proposed work was applied to four datasets, namely D1, D2, D3, and D4. D1 was collected from the UCI repository [46], while D2 was collected from Mendeley [47] and contains 48 features. D3 and D4 were also collected from Mendeley [48], with D3 containing 111 features and 58,645 instances, and D4 containing 111 features and 88,647 instances. Each dataset consists of two classes: phishing and legitimate. A detailed description of each dataset is provided in Table 1.

It should be noted that the UCI phishing dataset [46] and two variants of Mendeley [48] are imbalanced. As discussed in Section III-C, we applied a data re-sampling method to balance the datasets and improve the performance of our proposed model.

B. HYBRID FEATURE SELECTION APPROACH

The proposed approach for feature selection in this research involves the use of three boosting models, namely XGB, CatB, and LGBM. These models are used to extract essential features from the datasets under consideration. The feature importance for all the features is then computed separately using XGB, CatB, and LGBM. The average feature importance is then taken for each feature across the three selected feature subsets based on their feature importance scores as shown in the following equation. $AVG_{F_i} = \frac{1}{n} \sum_{j=1}^{n} RF_i^j$,

TABLE 2. The performance of Five Boosting classifiers & BMLSELM algorithm with all datasets with all features.

	Measures	LGBM	XGB	CatB	AdaB	GB	BMLSELM
With Dataset **D1**	TPR	96.74	97.32	97.7	97.87	95.95	98.39
	TNR	97.65	97.83	97.51	96.64	94.93	97.45
	Precision	97.22	97.42	97.02	95.94	93.86	96.93
	F-Score	96.98	97.37	97.36	96.9	94.89	97.65
	MCC	94.44	95.17	95.16	94.35	90.7	95.72
	Accuracy	97.24	97.6	97.6	97.19	95.38	97.87
With Dataset **D2**	TPR	98.7	98.8	98.6	97.67	98.1	98.61
	TNR	98.89	98.99	98.79	96.34	97.9	99.29
	Precision	98.9	99	98.8	96.3	97.9	99.3
	F-Score	98.8	98.9	98.7	96.98	98	98.95
	MCC	97.6	97.8	97.4	94	96	97.9
	Accuracy	98.8	98.9	98.7	97	98	98.95
With Dataset **D3**	TPR	94.89	95.68	95.32	92.63	92.99	95.91
	TNR	95.48	96.03	96.08	94.57	93.31	96.38
	Precision	94.93	95.54	95.61	93.99	92.45	95.94
	F-Score	94.91	95.61	95.47	93.3	92.72	95.92
	MCC	90.38	91.71	91.43	87.27	86.27	92.3
	Accuracy	95.2	95.87	95.72	93.64	93.16	96.16
With Dataset **D4**	TPR	97.45	97.75	97.77	97.69	96.64	98.14
	TNR	97.23	97.68	97.61	97.18	95.21	98.11
	Precision	97.23	97.69	97.62	97.17	95.15	98.12
	F-Score	97.34	97.72	97.7	97.43	95.89	98.13
	MCC	94.68	95.34	95.39	97.88	91.84	96.26
	Accuracy	97.34	97.71	97.69	97.43	95.91	98.13

Where AVG_{F_i} is an average of i^{th} feature importance (where $i = [1, m]$) when there are m features in a respective dataset, $n = 3$ (since, we employed three models such as XGB, CatB, and LGBM to obtain feature importance of each feature), RF_i^j is an importance of i^{th} feature of j^{th} model. This approach ensures that the most important features are selected, as they have highest score across all three boosting models.

After the average feature importance is computed, the K-topmost feature subset is selected to obtain the highest accuracy. The K-topmost feature subset is the set of features with the highest ranking and is chosen based on their relative importance. This hybrid approach helps to improve the accuracy of the proposed model by selecting only the most important features.

The stepwise approach for the selection of K-topmost features presented in Figure 5, and table 1 provides the optimal percentage of features selected from each phishing dataset, along with the relevant number of features based on the selected percentage. This approach ensures that the most relevant features are retained while minimizing the risk of overfitting.

C. DATA BALANCING

Imbalanced datasets can be addressed using data balancing techniques such as Random Under Sampling (RUS) and Random Over Sampling (ROS) [50]. In this study, we apply data balancing techniques to the K-topmost selected feature subsets of three datasets, namely D1, D3, and D4, which initially had imbalanced data.

For instance, the D1 dataset has 4898 legitimate instances and 6157 phishing instances. To balance this dataset, we use the ROS method, which randomly duplicates the instances of the minority class (legitimate in this case) and adds them to itself until the number of instances in the minority class is equal to the majority class (phishing in this case). This results in a balanced dataset with a total of 12314 instances, where each class has 6157 instances.

Similarly, in the D3 dataset, the phishing class is the minority class with 27998 instances, while the legitimate class is the majority class with 30647 instances. Using the ROS method, we duplicate the phishing class instances until we have 30647 instances, resulting in a balanced dataset with a total of 61294 instances.

Finally, in the D4 dataset, the legitimate class with 30647 instances is the minority class, while the phishing class with 58000 instances is the majority class. We duplicate the legitimate class instances until we have 58000 instances, resulting in a balanced dataset with a total of 116000 instances.

It is worth noting that the data balancing step was necessary to ensure that our models were trained on a balanced dataset,

TABLE 3. The performance of Five Boosting classifiers & BMLSELM algorithm with D1 using 20 features.

	Measures	LGBM	XGB	CatB	AdaB	GB	BMLSELM
Without Data Balance	TPR	96.64	97.61	97.6	97.39	96.13	97.9
	TNR	97.33	97.59	97.43	96.78	94.78	97.6
	Precision	96.83	97.21	96.93	96.13	93.66	97.12
	F-Score	96.73	97.36	97.26	96.76	94.88	97.51
	MCC	93.98	95.16	94.98	94.07	90.71	95.44
	Accuracy	97.01	97.6	97.51	97.06	95.38	97.73
With Data Balance	TPR	97	98.04	97.95	98.44	95.38	98.43
	TNR	96.33	96.76	96.44	97.01	94.78	96.38
	Precision	96.37	96.78	96.46	97.02	94.85	96.37
	F-Score	96.69	97.4	97.2	97.73	95.11	97.39
	MCC	93.34	94.81	94.4	95.46	90.17	94.82
	Accuracy	96.67	97.4	97.19	97.72	95.08	97.4

TABLE 4. The performance of Five Boosting classifiers & BMLSELM algorithm with D2 using 33 features.

	Measures	LGBM	XGB	CatB	AdaB	GB	BMLSELM
With Data Balance	TPR	98.8	98.51	98.6	97.97	98.1	98.7
	TNR	98.79	99.09	98.59	96.54	97.9	98.89
	Precision	98.8	99.1	98.6	96.5	97.9	98.9
	F-Score	98.8	98.8	98.6	97.23	98	98.8
	MCC	97.59	97.6	97.19	94.51	96	97.6
	Accuracy	98.8	98.8	98.6	97.25	98	98.8

TABLE 5. The performance of Five Boosting classifiers & BMLSELM algorithm with D3 using 96 features.

	Measures	LGBM	XGB	CatB	AdaB	GB	BMLSELM
Without Data Balance	TPR	94.89	95.34	95.24	94.48	93.06	95.9
	TNR	95.48	96.1	96.17	95.12	93.34	96.21
	Precision	94.93	95.63	95.72	94.53	92.48	95.74
	F-Score	94.91	95.49	95.48	94.5	92.77	95.82
	MCC	90.38	91.46	91.44	89.61	86.37	92.22
	Accuracy	95.2	95.74	95.73	94.82	93.21	96.06
With Data Balance	TPR	95.04	95.73	95.59	94.54	93.46	95.8
	TNR	94.73	95.69	95.31	94.22	92.98	96.57
	Precision	94.86	95.81	95.42	94.36	93.13	96.69
	F-Score	94.95	95.77	95.5	94.45	93.3	96.24
	MCC	89.78	91.43	90.91	88.77	86.45	92.36
	Accuracy	94.89	95.71	95.45	94.38	93.22	96.18

which can improve their performance in detecting phishing attacks.

D. BMLSELM

The MLSELM based on boosting techniques utilized four boosting models, namely XGB, CatB, LGBM, and AdaB, out of the five available, as GB's performance was inadequate. Its architecture includes three layers, as depicted in Figure 4. The first layer integrates all four boosting models, while the second layer integrates three models except for AdaBoost. The last layer employs XGB as the meta-learner. Four phishing datasets, containing all features, were used as input to the BMLSELM and the five boosting models in the first phase, followed by the evaluation of the unbalanced K-topmost selected features of each dataset through BMLSELM and the five boosting models in the second phase. Finally, the balanced K-topmost selected features of each dataset were evaluated through BMLSELM as shown in Figure 5. The proposed model is designed to achieve significant detection rate using hybrid feature selection. It is evaluated on different datasets to evaluate the behavior of the model with varying datasets. The model can be deployed as a web application or

TABLE 6. The performance of Five Boosting classifiers & BMLSELM algorithm with D4 using 83 features.

	Measures	LGBM	XGB	CatB	AdaB	GB	BMLSELM
Without Data Balance	TPR	97.2	97.48	97.63	97.45	96.39	97.82
	TNR	95.08	95.4	95.92	95.07	93.01	96.42
	Precision	97.37	97.54	97.82	97.35	96.24	98.09
	F-Score	97.29	97.51	97.72	97.4	96.32	97.95
	MCC	92.22	92.85	93.47	92.57	89.46	94.13
	Accuracy	96.46	96.75	97.03	96.62	95.21	97.33
With Data Balance	TPR	97.24	97.72	97.58	97.8	96.5	97.98
	TNR	96.72	97.3	97.3	97.37	94.62	97.78
	Precision	96.73	97.31	97.32	97.38	94.56	97.8
	F-Score	96.77	97.51	97.45	97.59	95.52	97.89
	MCC	93.97	95.02	94.89	95.17	91.12	95.77
	Accuracy	96.98	97.51	97.44	97.58	95.55	97.88

TABLE 7. The performance BMLSELM with hybrid feature selection on datasets.

	Measures	D1(20)	D2(33)	D3(96)	D4(83)
BMLSELM Unbalanced	Recall	97.9	98.70	95.90	97.82
	Precision	97.12	98.90	95.74	98.09
	F-Score	97.51	98.80	95.82	97.95
	Accuracy	97.73	98.80	96.06	97.33
BMLSELM Balanced	Recall	98.43	98.70	95.80	97.98
	Precision	96.37	98.90	96.69	97.8
	F-Score	97.39	98.80	96.24	97.89
	Accuracy	97.4	98.80	96.18	97.88

TABLE 8. The performance of MLSELM and BMLSELM with all datasets with all features.

	Measures	D1	D2	D3	D4
MLSELM Balanced	Recall	98.07	99.28	96.70	98.96
	Precision	97.34	98.48	96.84	97.93
	F-Score	97.70	98.88	96.77	98.44
	Accuracy	97.76	98.90	96.79	98.43
MLSELM Unbalanced	Recall	98.05	99.28	96.25	98.16
	Precision	95.08	98.48	96.42	97.88
	F-Score	96.54	98.88	96.33	98.02
	Accuracy	97.06	98.90	96.5	97.41
BMLSELM Balanced	Recall	97.18	98.61	95.70	98.14
	Precision	97.01	99.30	96.61	98.12
	F-Score	97.10	98.95	96.15	98.13
	Accuracy	97.15	98.95	96.13	**98.13**
BMLSELM Unbalanced	Recall	98.39	98.61	95.91	97.83
	Precision	96.93	99.30	95.94	98.07
	F-Score	97.65	98.95	95.92	97.95
	Accuracy	**97.87**	**98.95**	**96.16**	97.33

a browser extension which takes input as URL and source code of the websites and can result the web page as either legitimate or phishing.

IV. EXPERIMENTATION RESULTS

In this study, we applied the proposed BMLSELM algorithm with five boosting based Machine Learning algorithms, including CatB, LGBM, GB, AdaB, and XGB, to four datasets listed in Table 1. The classification metrics used to evaluate the performance of the models include Precision, Recall, F-score, and Accuracy. In this study, we considered phishing instances as positive and legitimate instances as negative. The calculation of each metric was based on the following definitions:

- P: Indicates total count of phishing instances
- N: Indicates total count of legitimate instances
- T_N:The predicted count of legitimate instances that are correctly classified as legitimate by the model.
- F_N: The predicted count of phishing instances that are incorrectly classified as legitimate by the model.
- T_P:The predicted count of phishing instances that are correctly classified as phishing by the model.
- F_P:The predicted count of legitimate instances that are incorrectly classified as legitimate by the model.

The calculation of each metric is shown below:

- Precision=$\frac{T_P}{T_P+F_P} \times 100$
- Recall=$\frac{T_P}{T_P+F_N} \times 100$
- F-score=$\frac{\text{precision}\times\text{Recall}}{\text{precision}+\text{Recall}} \times 100$
- Accuracy=$\frac{T_P+T_N}{P+N} \times 100$

We evaluated the performance of BMLSELM and compared it with five classification models on four datasets (D1,D2,D3, and D4) with all features, as well as on balanced and unbalanced K-topmost features, as described in section IV-C. Additionally, we conducted a comparative analysis of the results of BMLSELM on four phishing datasets with the existing literature, which is presented in section IV-D.

A. EXPERIMENT 1: EVALUATION OF BMLSELM ACROSS ALL DATASETS WITHOUT FEATURE SELECTION

In this section, we experimented various boosting algorithms applied on all features in D1, D2, D3 and D4 datasets. The results of individual boosting algorithms are

TABLE 9. The Performance Comparison of BMLSELM with MLSELM and existing literature which employed D2 phishing dataset.

Method	FS method	Features	Proposed Model	Accuracy	Precision	Recall	F1-Score
[32]	HEFS: hybrid perturbation ensemble and CDF-g.	10	RF	94.6	–	–	–
[15]	IG, ReliefF	20	RF	98.11	–	–	–
[29]		10	SC+REF and RF	97.6	–	–	–
[16]	Chi-Square ,IG,GR,PCC,PCA	32	IG+RF	98.38	–	–	–
[23]	Consensus	17		98.17	97.9	98.37	98.13
	majority voting	23		98.63	98.48	98.81	98.64
[36]	FRS	30	RF	–	–	–	95
[34]	LAPPSO	17	LAPPSO+DT	–	–	–	95.8
	FBPSO	26	FBPSO+DT	–	–	–	85.9
[17]	Chi-Squared	28	RF	97.8	97.5	97.8	97.8
	KMO	34	RF	–	–	–	–
	REF	26	RF	–	–	–	–
	Pearson Correlation	26	RF	–	–	–	–
[9]	IG, ANOVA, RFE, ReliefF, and Fisher Score	–	Ensemble IG+ANOVA+REF followed by Ensemble of XGBoost, DT, and RF.	98.45	–	–	–
MLSELM [45]	Nil	48	MLSELM	98.90	98.48	99.28	98.88
Proposed BMLSELM	Without Feature selection	48	BMLSELM	**98.95**	99.30	98.61	98.95
Proposed BMLSELM	Hybrid Feature selection	33	BMLSELM	**98.80**	98.9	98.7	98.8

also compared with proposed BMLSELM which can be seen in Table 2. From the results, it is observed that XGB outperformed other boosting algorithms across all datasets. Also, the results demonstrate that the proposed BMLSELM has achieved significant performance in accuracy and MCC across all datasets compared to XGBoost algorithm.

B. EXPERIMENT 2: EVALUATION OF BMLSELM ACROSS ALL DATASETS WITH FEATURE SELECTION

In this section, we apply feature selection prior to the model training and dataset with the selected features are fed to the proposed model for the classification. The embedding method with features selected from boosting algorithms through feature importance is applied across all datasets. The top k features from the boosting algorithms are choosen for the final features selection. From the experimental analysis, k is chosen as 20 for D1, 33 for D2, 96 for D3, 83 for D4 datasets. The results with boosting algorithms and BMLSELM on D1 dataset is shown in Table 3. From the results, it is clearly seen that BMLSELM outperformed other boosting algorithms with an accuracy and MCC of 97.73 and 95.44 with imbalanced data. Also, the proposed model performed better than other boosting algorithms when balanced data is fed to the model. But, the proposed model did perform well when the imbalanced data is given as input compared to the balanced data. Note that, these results from the proposed model includes only 20 features from balanced and imbalanced data.

As D2 is already balanced, we conducted the experiment with feature selection on the balanced data. The results with proposed model and other classifiers is shown in Table 4. From the results, it is demonstrated that the proposed model BMLSELM with 33 selected features achieved significant performance with an accuracy of 98.8 and MCC of 97.6. It is also observed that XGB achieved the similar performance compared to BMLSELM but with slightly lower in TPR.

Similarly, the traditional boosting algorithms and the proposed model is applied on D3 and D4 datasets with 96 and 83 selected features respectively. The results with D3 dataset is shown in Table 5. From the results, it is observed that BMLMSELM performed better with and without data balance compared to other boosting algorithms with an accuracy of 96.18%, MCC of 92.36. The results with D4 dataset is given in Table 6. From the results, it is observed that BMLSELM achieved an accuracy of 97.33 and MCC of 94.13 with imbalanced data and an accuracy of 97.88 and MCC of 95.77 with balanced data.

C. THE COMPARISON OF BOOSTING ALGORITHMS AND MLSELM WITH BMLSELM

In this section, we compare our proposed work with our existing work MLSELM as they are experimented on same datasets and use stacking mechanism. The comparison results are shown in Table 8 and 7. From the results, it is observed that, MLSELM on D1 dataset achieved an accuracy of 97.76 with balanced data and 97.06 with imbalanced data whereas the proposed model BMLSELM achieved

TABLE 10. The Performance Comparison of BMLSELM with MLSELM and existing literature which employed D1 phishing dataset.

Method	FS method	Features	Proposed Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
[31]	IG, GR, Relief-F, REF, and PCA	27	Stacking (NN+RF+Bagging)	97.4	96	98.1	97
[19]	OFS	26	NN	96.75	96.65	97.46	97.05
[24]	ReliefF	17	Majority voting with multiple models	95	95	–	94
[21]	FSOR and	22	FSOR+RF	97.18	–	–	–
	FSFM	9	FSFM+RF	95.21	–	–	–
[25]	WSER	28	RF	97.29	97	97.3	–
[30]	GSA,CFS,IG and PCA	15	GSA+RF	95.53	–	–	–
[33]	GA+DT	20	NR-HS	92.8	–	–	–
[23]	Consensus	9	LGBM+AdaBoost	93.55	92.65	95.82	94.21
	majority voting	13	LGBM+AdaBoost	95.29	95.23	96.63	95.92
[36]	FRS	24	RF	–	–	–	93
[34]	LAPPSO	20	LAPPSO+DT	–	–	–	96.6
	FBPSO	18	FBPSO+DT	–	–	–	89.2
[17]	Chi-Squared	15	RF	96.2	95.7	97.5	96
	KMO	19	RF	–	–	–	–
	REF	15	RF	–	–	–	–
	Pearson Correlation	14	RF	–	–	–	–
[9]	IG, ANOVA, RFE, ReliefF, and Fisher Score	–	Ensemble IG+ANOVA+REF followed by Ensemble of XGBoost, DT, and RF.	97.51	97.13	–	97.27
[18]	CFS	23	KNN	97	96	97	97
[20]	FVV index	23	NN	94.5	96.4	93.6	95
[26]	AR	26	RF	97.31	–	–	–
[22]	RFR	9	RF	95.4	93.5	95.9	94.7
MLSELM [45] with data balanced	Nil	30	MLSELM	97.76	97.34	98.07	97.7
MLSELM [45] without data balanced	Nil	30	MLSELM	97.06	95.08	98.05	96.54
Proposed BMLSELM	Without Feature selection	30	BMLSELM	**97.87**	96.93	98.39	97.65
BMLSELM	Hybrid Feature selection	20	BMLSELM	**97.73**	97.12	97.9	97.51

TABLE 11. The Performance Comparison of BMLSELM with MLSELM and existing literature which employed D4 phishing dataset.

Method	FS method	Features	Proposed Model	Accuracy	Precision	Recall	F1-Score
[27]	RF based feature importance	14	RF	97	–	–	–
[37]	DEFSTH	40	NB	96.82	–	–	95.38
[35]	BSSA	49	BSSA-X+KNN	95.07	–	–	–
MLSELM [45] with unbalanced data	Nil	111	MLSELM	97.41	97.88	98.16	98.02
MLSELM [45] with balanced data	Nil	111	MLSELM	98.43	97.93	98.96	98.44
BMLSELM with unbalanced data	without feature selection	111	BMLSELM	97.33	98.07	97.83	97.95
BMLSELM with balanced data	without feature selection	111	BMLSELM	98.13	98.12	98.14	98.13
BMLSELM with unbalanced data	Hybrid feature selection	83	BMLSELM	97.33	98.09	97.82	97.95
BMLSELM with balanced data	Hybrid feature selection	83	BMLSELM	97.88	97.8	97.98	97.89

TABLE 12. The Performance Comparison of BMLSELM with existing literature which employed D3 phishing dataset.

Method	FS method	Features	Proposed Model	Accuracy	Precision	Recall	F1-Score
MLSELM [45] with unbalanced data	Nil	111	MLSELM	96.50	96.42	96.25	96.33
MLSELM [45] with balanced data	Nil	111	MLSELM	96.79	96.84	96.70	96.77
BMLSELM with unbalanced data	without feature selection	111	BMLSELM	96.16	95.94	95.91	95.92
BMLSELM with balanced data	without feature selection	111	BMLSELM	96.13	96.61	95.70	96.15
BMLSELM with unbalanced data	Hybrid feature selection	96	BMLSELM	96.06	95.74	95.90	95.82
BMLSELM with balanced data	Hybrid feature selection	96	BMLSELM	96.18	96.69	95.80	96.24

better accuracy of 97.87 without feature selection and an accuracy of 97.73% with only 20 selected features. Moreover, BMLSELM outperformed MLSELM with an accuracy of 98.95 with all features and even had achieved significant performance of 98.80 accuracy with only 33 features.

However, the MLSELM model performed slightly better than BMLSELM on the D3 and D4 datasets. On D3, the MLSELM model achieved an accuracy of 96.79% and 96.5% under balanced and unbalanced categories, respectively, with all features, while BMLSELM achieved 96.13% and 96.16% under balanced and unbalanced categories, respectively. The performance difference between the two models on D3 was only 0.66% and 0.34%, respectively. However, the proposed model on D3 with 96 features achieved a significant performance with an accuracy of 96.18% on balanced data and 96.06% with unbalanced data.

Similarly, on D4, MLSELM achieved an accuracy of 98.43% and 97.41% under balanced and unbalanced categories, respectively, with all features, while BMLSELM achieved 98.13% and 97.33% under balanced and unbalanced categories, respectively. The performance difference between the two models on D4 was negligible at 0.3% and 0.08%, respectively. However, BMLSELM with 83 reduced features achieved a significant performance with an accuracy of 97.88% on balanced data and 97.33% with unbalanced data.

D. THE COMPARISON OF BMLSELM WITH EXISTING LITERATURE

In this section, we compare various existing works with our proposed work that used same datasets for their experimentation. The comparison results with D1 dataset is given in Table 10. From the table, it is clearly visible that the proposed model achieved better performance than existing works with an accuracy of 97.87 with all features and 97.73% with only 20 features. The second comparison results with D2 dataset is shown in Table 9. From the table, it is demonstrated that BMLSELM outperformed existing works with an accuracy of 98.80 with feature selection and 98.95 with all features. On D3 MLSELM obtained 96.79% with 111 features where as BMLSELM achieved 96.16% of accuracy with 96 features where the difference is 0.63% only. Finally, the comparison results with D4 dataset in Table 11 shows that BMLSELM performed lower compared to our earlier work but it has achieved significant performance of accuracy 97.88% with only 83 features compared to 98.43% with 111 features.

V. CONCLUSION

In this paper, we proposed a feature selection based stacking model (BMLSELM)that uses various boosting algorithms to identify relevant features. Also, the boosting algorithms are used to generate multi stacking model with estimators at different layers to achieve significant performance. BMLSELM is applied on D1,D2,D3 and D4 datasets to evaluate the performance of the model across different datasets. The model achieved significant performance with D1 to D4 datasets in two cases i.e. datasets with feature selection and without feature selection. The model is experimented with both balanced and imbalanced data. The experimental results of BMLSELM with D1-D4 datasets demonstrates that the model achieved an significant accuracy of 97.4 (D1 with 20 features), 98.80(D2 with 33 features), 96.18(D3 with 96 features) and 97.88(D4 with 83 features). Finally, the model is compared with baseline models where it outperformed the existing models with significant difference across different metrics. In the future work, we would like to use different feature selection ensembles, clustering algorithms and feature engineering techniques for the hidden feature generation that helps in improving the detection accuracy of the model.

REFERENCES

[1] 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.

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

[3] 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*, Jun. 2009, pp. 1245–1254.

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

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

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

[7] 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, Oct. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X11000045

[8] 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.

[9] 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.

[10] 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.

[11] 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?doid=2019599.2019606

[12] 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.

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

[14] 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. (ICOIACT)*, Jul. 2019, pp. 578–583.

[15] A. Abuzuraiq, M. Alkasassbeh, and M. Almseidin, "Intelligent methods for accurately detecting phishing websites," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2020, pp. 85–90.

[16] Md. S. M. Prince, A. Hasan, and F. M. Shah, "A new ensemble model for phishing detection based on hybrid cumulative feature selection," in *Proc. IEEE 11th IEEE Symp. Comput. Appl. Ind. Electron. (ISCAIE)*, Apr. 2021, pp. 7–12.

[17] 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. (CONECCT)*, Jul. 2020, pp. 1–6.

[18] S. Mohanty and A. A. Acharya, "MFBFST: Building a stable ensemble learning model using multivariate filter-based feature selection technique for detection of suspicious URL," *Proc. Comput. Sci.*, vol. 218, pp. 1668–1681, Jan. 2023.

[19] 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 Int. Conf. Parallel Distrib. Process. With Appl., Ubiquitous Comput. Commun., Big Data Cloud Comput., Social Comput. Netw., Sustain. Comput. Commun. (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, Dec. 2018, pp. 781–787.

[20] 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.

[21] S. Shabudin, N. Samsiah, K. Akram, and M. Aliff, "Feature selection for phishing website classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 1–9, 2020.

[22] A. A. Ubing, 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.

[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] 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.

[25] M. D. Abdulrahman, J. K. Alhassan, O. S. Adebayo, J. A. Ojeniyi, and M. Olalere, "Phishing attack detection based on random forest with wrapper feature selection method," *Int. J. Inf. Process. Commun.*, vol. 7, pp. 209–224, 2019.

[26] M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," in *Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS)*, Mar. 2018, pp. 1–5.

[27] Y. Wei and Y. Sekiya, "Feature selection approach for phishing detection based on machine learning," in *Proc. Int. Conf. Appl. CyberSecurity (ACS)*. Cham, Switzerland: Springer, 2022, pp. 61–70.

[28] M. Almseidin, A. A. Zuraiq, M. Al-Kasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," Int. Assoc. Online Eng., 2019.

[29] J. Moedjahedy, A. Setyanto, F. K. Alarfaj, and M. Alreshoodi, "CCrFS: Combine correlation features selection for detecting phishing websites using machine learning," *Future Internet*, vol. 14, no. 8, p. 229, Jul. 2022.

[30] 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.

[31] 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, 2020.

[32] K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci.*, vol. 484, pp. 153–166, May 2019.

[33] 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.

[34] M. Akhavan and S. M. Hossein Hasheminejad, "An unsupervised feature selection for web phishing data using an evolutionary approach," in *Proc. 7th Int. Conf. Web Res. (ICWR)*, May 2021, pp. 41–47.

[35] R. A. Khurma, K. E. Sabri, P. A. Castillo, and I. Aljarah, "Salp swarm optimization search based feature selection for enhanced phishing websites detection," in *Proc. 24th Int. Conf. Appl. Evol. Comput. (EvoApplications)*. Cham, Switzerland: Springer, Apr. 2021, pp. 146–161.

[36] M. Zabihimayvan and D. Doran, "Fuzzy rough set feature selection to enhance phishing attack detection," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jun. 2019, pp. 1–6.

[37] L. Brezočnik, I. Fister, and G. Vrbančič, "Applying differential evolution with threshold mechanism for feature selection on a phishing websites classification," in *Proc. Workshops BBIGAP, QAUCA, SemBDM, SIM-PDA, M2P, MADEISD, Doctoral Consortium*, Bled, Slovenia. Cham, Switzerland: Springer, Sep. 2019, pp. 11–18.

[38] S. Khan, M. Khan, N. Iqbal, M. Li, and D. M. Khan, "Spark-based parallel deep neural network model for classification of large scale RNAs into piRNAs and non-piRNAs," *IEEE Access*, vol. 8, pp. 136978–136991, 2020.

[39] F. Khan, M. Khan, N. Iqbal, S. Khan, D. Muhammad Khan, A. Khan, and D.-Q. Wei, "Prediction of recombination spots using novel hybrid feature extraction method via deep learning approach," *Frontiers Genet.*, vol. 11, Sep. 2020, Art. no. 539227.

[40] P. Vaitkevicius and V. Marcinkevicius, "Composition of ensembles of recurrent neural networks for phishing websites detection," in *Proc. Int. Baltic Conf. Databases Inf. Syst.* Cham, Switzerland: Springer, 2020, pp. 297–310.

[41] 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.

[42] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15196–15209, 2019.

[43] A. Mughaid, S. AlZu'bi, A. Hnaif, S. Taamneh, A. Alnajjar, and E. A. Elsoud, "An intelligent cyber security phishing detection system using deep learning techniques," *Cluster Comput.*, vol. 25, no. 6, pp. 3819–3828, Dec. 2022.

[44] M. A. Adebowale, K. T. Lwin, and M. A. Hossain, "Intelligent phishing detection scheme using deep learning algorithms," *J. Enterprise Inf. Manag.*, vol. 36, no. 3, pp. 747–766, Apr. 2023.

[45] L. R. Kalabarige, R. S. Rao, A. Abraham, and L. A. Gabralla, "Multilayer stacked ensemble learning model to detect phishing websites," *IEEE Access*, vol. 10, pp. 79543–79552, 2022.

[46] R. M. Mohammad, F. Thabtah, and L. McCluskey, "UCI machine learning repository," UCI Repository, Tech. Rep., 2015.

[47] C. L. Tan, "Phishing dataset for machine learning: Feature evaluation," Mendeley Data, 2018, vol. 1, p. 2018.
[48] G. Vrbancic, "Phishing websites dataset," Mendeley Data, 2020, vol. 1.
[49] S. Dangwal and A. Moldovan, "Feature selection for machine learning-based phishing websites detection," in *Proc. Int. Conf. Cyber Situational Awareness, Data Anal. Assessment (CyberSA)*, Jun. 2021, pp. 1–6.
[50] L. R. Kalabarige and H. Maringanti, "Symptom based COVID-19 test recommendation system using machine learning technique," *Intell. Decis. Technol.*, vol. 16, no. 1, pp. 1–11, 2022.

LAKSHMANA RAO KALABARIGE received the Ph.D. degree in wireless and cognitive radio networks from Gitam University, Visakhapatnam. He is currently an Associate Professor with 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 cyber security from NITK Surathkal. He is currently an Associate Professor with the GITAM School of Technology, Visakhapatnam, GITAM (Deemed to be University), India. His research interests include information security, cyber security, phishing, machine learning, and natural language processing.

ALWYN R. PAIS received the B.Tech. degree in CSE from Mangalore University, India, the M.Tech. degree in CSE from IIT Bombay, India, and the Ph.D. degree from NITK, India. He is an Associate Professor with the Department of Computer Science and Engineering, NITK Surathkal, India. His interests include information security, image processing, and computer vision.

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 was a Senior Fellow of 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.