I. J. Computer Network and Information Security, **2022, 6, 37-53**
Published Online on December 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2022.06.04

Performance Evaluation of Machine Learning-based Robocalls Detection Models in Telephony Networks

Bodunde O. Akinyemi*
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
E-mail: bakinyemi@oauife.edu.ng
ORCID iD: https://orcid.org/0000-0001-6943-7137
*Corresponding Author

Oluwatoyin H. Odukoya
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
E-mail: olagun12@oauife.edu.ng
ORCID iD: https://orcid.org/0000-0003-2894-7231

Mistura L. Sanni
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
E-mail: msanni@oauife.edu.ng
ORCID iD: https://orcid.org/0000-0001-9206-4009

Gilbert Sewagnon
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
E-mail: gsewagnon@oauife.edu.ng
ORCID iD: https://orcid.org/0000-0002-7035-8847

Ganiyu A. Aderounmu
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
E-mail: gaderoun@oauife.edu.ng
ORCID iD: https://orcid.org/0000-0002-7992-514X

Received: 12 February 2022; Revised: 08 June 2022; Accepted: 27 August 2022; Published: 08 December 2022

Abstract: Many techniques have been proposed to detect and prevent spam over Internet telephony. Human spam calls can be detected more accurately with these techniques. However, robocalls, a type of voice spammer whose calling patterns are similar to those of legitimate users, cannot be detected as effectively. This paper proposes a model for robocall detection using a machine learning approach. Voice data recordings were collected and the relevant features for study were selected. The selected features were then used to formulate six (6) detection models. The formulated models were simulated and evaluated using some performance metrics to ascertain the model with the best performance. The C4.5 decision tree algorithm gave the best evaluation result with an accuracy of 99.15%, a sensitivity of 0.991%, a false alarm rate of 0.009%, and a precision of 0.992%. As a result, it was concluded that this approach can be used to detect and filter both machine-initiated and human-initiated spam calls.

Index Terms: Spam, VoIP, Robocalls, SPIT, Machine Learning.

1. Introduction

Voice over Internet Protocols (VoIPs) are becoming increasingly popular, and Spam over IP Telephony (SPIT) is expected to become an increasingly serious issue in the near future. These nuisance calls are considered far more bothersome than messages or social spam since they are produced in real-time and have a much greater impact on productivity [1]. There exists a means of transmitting unsolicited calls over the traditional Public Switched Telephone

Network (PSTN), where such calls are commonly initiated by telemarketers. However, the high cost of PSTN calls in comparison to email or VoIP communications tends to make this form of advertising less attractive to telemarketers. The costs a spammer would incur if Internet telephony is used, on the other hand, are significantly lower. The cost of sending SPIT is about three orders of magnitude less than traditional PSTN telemarketer calls. Thus, the pricing structure of Internet telephony remains a significant contributor to its growth [1].

Telemarketing is the most common attack that occurs in VoIP environments. In the United States, the average person receives two to three telemarketing calls per day [2]. To sell their services or products, telemarketers target people in a specific industry or domain. Numerous companies have been able to open low-cost off-shore offices to conduct these types of telemarketing campaigns with VoIP technology. SPIT are classified as human spammers. A third type of spam call that is extremely difficult to stop are those that are initiated by call bots, which are automated machines [3]. They are similar to telemarketers in that calls are generated by computers or pre-recorded machines. They use a pre-defined list of numbers, and the message is typically the same throughout. Robocalls and botnets are two terms for the same phenomenon [4].

A "robocall," also known as "voice broadcasting," is any phone call that sends a prerecorded message to a large number of people via an automatic (computerized) telephone dialing system, also known as an automatic dialer or "autodialer." Robocalls are popular in many industries, including real estate, telemarketing, and direct sales. The Telephone Consumer Protection Act (TCPA) of 1991 stated that the vast majority of businesses that use robocalling are legitimate, but some are not. Those shady businesses may be more than just annoyances to customers; they could also be attempting to defraud them. The use of robocalls allows organizations to reach thousands of potential customers in a short amount of time, unlike traditional "live" telemarketing methods [5].

Any prerecorded message can be used in a robocall. Many organizations are using this technology to offer specific products or services, send out hundreds of calls at a time, and track which ones may lead to business. Unfortunately, the nature of SPIT is different from email spam. Robocalls are becoming just as dangerous as phone calls in terms of social engineering. The term "robocall" refers to voice calls that are generated automatically. Although most people deliberately ignore robocalls due to their spam-like nature, they can sometimes appear legitimate if they claim to be from a known source. According to CNN Business, approximately 25 million Americans have been scammed by robocalls, with robocalls accounting for roughly half of all calls from unknown numbers [6, 7]. This is a threatening statistic, which is exacerbated by the fact that 90% of robocalls claim to know the recipient.

The content of SPIT cannot be verified before the callee is disturbed by the ringing phone, so common spam prevention measures cannot be applied effectively to SPIT. This usually results in end-user disruption, loss of time, money, and information [8,9]. Due to the similarity in call patterns between legitimate users and speech spammers, it is difficult to identify voice spammers. Effective robocall protection will be critical to the success of VoIP, especially on public networks and at the gateways between public networks and corporate networks.

Many techniques for detecting and preventing spam over Internet telephony have been proposed, but these approaches can only detect human spam calls, not robocalls. Usually, detection of robocalls is treated as a classification problem in which most existing solutions resolve this by manually engineering features of a phony phone call and then classifying them using state-of-the-art machine learning algorithms [10]. Machine learning-assisted analysis of robocalls in telecommunications can help network operators restore consumer confidence in their security. A computer spam detection model that can detect and filter both machine-initiated spam calls and human-initiated spam calls is needed, as is a way to better safeguard VoIP conversations from spammer behavior changes.

Thus, this study attempts to evaluate some classification models that employ machine learning techniques to accurately classify voice calls as robocalls or human calls. The study developed a model for detecting robocalls, carried out a comparative analysis between different machine learning models, and established the minimum number of features required to accurately detect robocalls. The model with the best performance was then selected for the detection of robocalls in telephony networks.

The remaining part of the paper is arranged as follows: The literature review is covered in section 2, while section 3 explains the methodology and presents the proposed architecture as well as the machine learning process. Section 4 discusses the simulation and evaluation results, while Section 5 concludes the paper.

2. Related Works

Over the past decade, researchers have been studying VoIP spam. Many techniques have been developed to combat spam, but none of them is capable of detecting all spam variants. As a rule, the techniques are variations of ways that have been developed in the field of e-mail spam [11,12]. This section examines some proposed techniques and their usefulness for VoIP spam prevention or mitigation.

With the rise in popularity of VoIP (Voice over Internet Protocol) systems, VoIP spam has exploded. Various methods based on the analysis of call behavior features have been proposed to effectively prevent spam calls [13]. A capability was introduced to determine whether a call is coming from the number it claims to be [14]. In [15], a model was proposed to control spam midway through a server based on Gray Listing and Black Listing recording spambanner based on the digits entered by the end client during the call. The spam table is used to analyses incoming calls. The outcome demonstrates that a single rundown is sufficient to filter every incoming call. If the call has a spam banner that is less than 10%, the call can enter; otherwise, the call will be a pre-recorded message. This system can attack all spam, but there is a slight loss of time in call arrangement, which is significantly less than the conventional filtering strategy.

In [16], a Hidden Markov Model was used to detect human and computer VoIP spammers in the Voice over IP protocol. In order to estimate spammer activity, the model is able to incorporate spammer behaviour from several sources into one model. It effectively distinguishes voice spammers from benign callers and avoids them before making the next call. As a result of the evaluation of the suggested model, it appears that it is more accurate in detecting human spam, although it does not detect all types of computer spam.

In the session initiation protocol, a model to detect unknown intrusions was created [17]. The technique allowed for the detection of unknown intrusions for which there were no signatures. A moderate change in the network environment can be accommodated by it automatically retraining itself. The actual throughput of this solution is not yet up to the standards of modern VoIP devices, but when used in conjunction with filtering mechanisms, it provides protection against huge traffic volumes.

In [18], a sensor system that uses distributed signature-based incorrect discovery to detect and prevent VoIP fraud was proposed. According to the assessment system's field tests, combining heterogeneous inputs and using a sophisticated processing technique improves detection accuracy, while a two-tier model executed by the intermediary handler located between the edge switch and the firewall was proposed [19], which is used to keep track of telemarketer calls and increase the recognition rate. The strategy established a link between the current parameter estimations and the parameter estimates of the new approaching call, as held by the dim rundown. This model exhibits a promising blocking probability rate of 0.90 against spammers, with an extremely low false positive rate of 0.01. However, detecting the spammer becomes impossible when the call comes from multiple sources.

According to the research work in [20], a behavior-based technique can be used to detect SPIT attacks. Network traces of voice communications and signals were gathered. There were nine SPIT attack identification criteria extracted from the traces, and a sliding window technique was implemented as a result of the findings. The investigation was limited by the fact that there were false positives and false negatives with the different classifiers used.

Different machine learning approaches have been used to detect SPIT attacks, e.g., the deep learning approach [21, 22]. In [23, 24], a new supervised learning-based technique was developed for detecting different sources of bots despite their botnet characteristics in a real-world, highly imbalanced network dataset. For both identification, they proposed three new generic IP-based features. The results showed that the system classifies different types of bots with an accuracy of 99.66%, which is the best result compared to previous work. However, the system often detects bots only after they have launched an attack. It was unable to detect a bot that was not successful in the attack stage, which was a major flaw in the software. Also, acoustic features of recorded voicemails to distinguish human calls from robocalls and spam calls from non-spam calls for human callers were presented [25]. In [26], a fully automated online Do-Not-Call registry list to avoid telemarketing calls and robocalls was implemented. The implementation of the registry has helped to reduce the number of lawful telemarketing calls. As a result of the Registry's efforts, unwanted telemarketing calls have been significantly reduced. However, not all robocalls can be avoided entirely. Despite being registered, certain consumers continue to get unwanted calls from third parties. In [27], spoofing was combated by proposing the "Do Not Originate" list and call authentication. The technology is effective in preventing unwanted robocalls by asking callers and carriers to confirm that they are the owners of a specific phone number. However, only spoofed calls from numbers on the "Do Not Originate" list are blocked by this technology.

In [28], a solution called "robocalls Mark and Sweep" was proposed, which has good performance in detecting and classifying SPIT. The system provides real-time scoring and classification of incoming calls based on predefined criteria such as: call duration, number validity, IP Black-and-white lists, calling patterns such as random and sequential dialing, destination number, acoustic fingerprinting, and Do Not Call list participation. The system gave a comprehensive resolution to handle SPIT at all points in the call process, involving post-call workflows like end user tagging. It provided the first-ever tool to identify "bad players" who are sourcing high SPIT volumes. The system itself alone could not adapt to new calling techniques. It was the first time any researcher had been able to identify "bad actors" who were sourcing large volumes of SPIT. The system as a whole was unable to adjust to new calling strategies.

Therefore, it was highlighted that the existing techniques can only detect human spam calls and not robocalls in view of the aforementioned approaches. A machine learning approach to spam detection has been proposed as one solution to this problem, e.g., the Ensemble Approach [22]. Previously proposed machine learning-based malicious call detection methods relied on various assumptions that a telephony network server could provide more information about the caller. Meanwhile, in [29, 30], a machine learning approach for telecommunications without making any assumptions about the underlying telephony networks was proposed, to prevent malicious calls over the telephone network. In [29], based on the proposed features, different state-of-the-art machine learning approaches were evaluated. There was a reduction in malicious calls of at least 90% and a precision of over 99.99% on legal calls. The proposed model, however, cannot effectively handle caller spoofing and cannot distinguish between spam and sales-related callers.

Therefore, machine learning models must be evaluated side-by-side in order to have the smallest number of features required for robocall detection to become more accurate.

3. Methodology

The use of supervised machine learning algorithms in the development of a predictive model for robocall detection using data collected from audio recordings of human calls and robocalls is the focus of this research. Figs. 1 and 2 depict the diagram of the methodological framework used to develop the predictive model and the SPIT detection architecture, respectively. The conceptual view of the model formulation process is shown in Fig. 3. The details of the process are presented as follows:

3.1. Data Collection and Preprocessing

Thousands of samples of Robocalls and human voices were acquired from data. world (https://data.world), an open data repository. Each voice sample collected was stored as a.WAV file, which is not a usable format for machine learning. The.WAV files were preprocessed for acoustic analysis using the specan function from the WarbleR R package. Filter-based and wrapper-based feature selection methods were applied to extract the most relevant and important features for Robocall detection. The reduced feature dataset was divided into training and testing datasets. An n-fold cross-validation evaluation was performed on both the training and testing datasets for each of the proposed supervised machine learning algorithms. The performance of each wrapper-based feature selection method in combination with the supervised machine learning algorithm was used to identify the most effective and efficient predictive model for robocall detection.

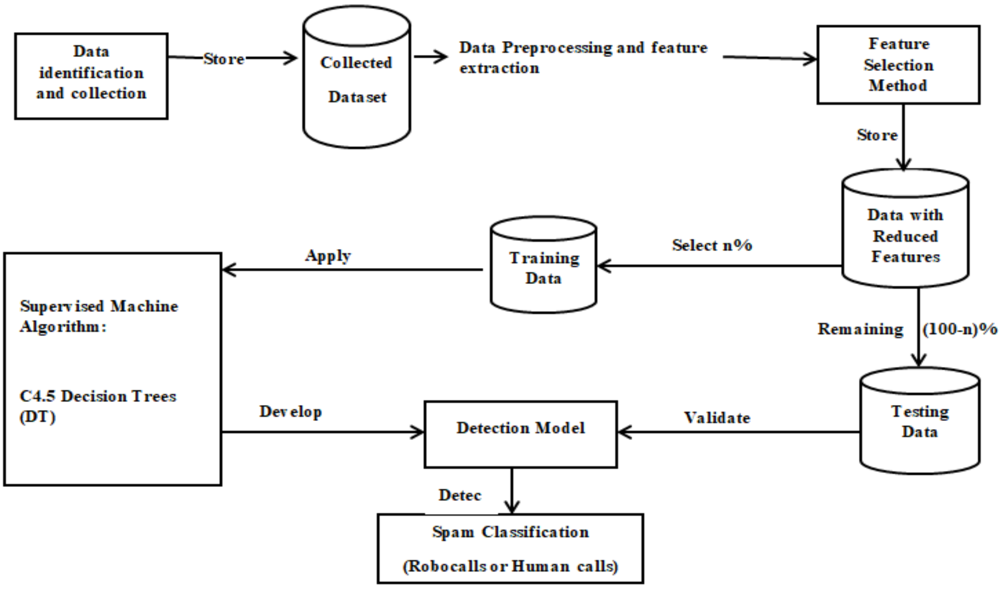


Fig.1. Robocalls Detection Framework

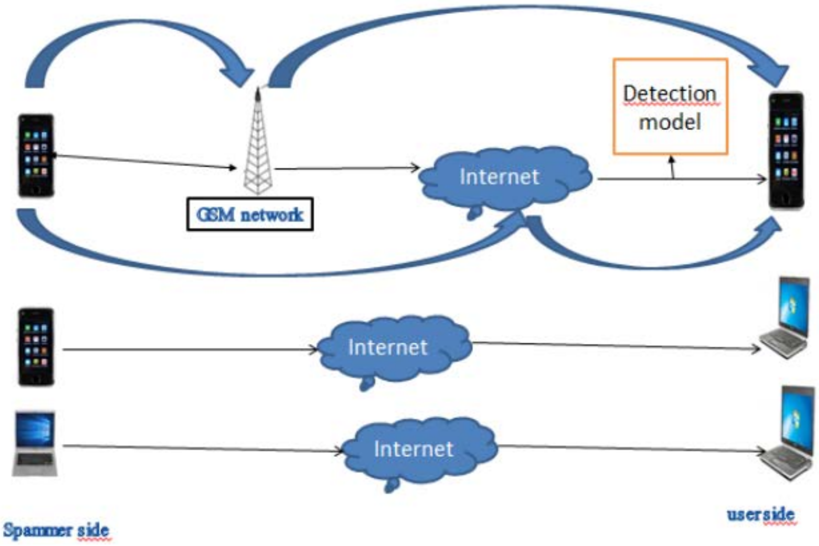


Fig.2. Proposed SPIT Detection Architecture

A. Wrapper- based Feature Selection Process

A simulation was carried out to evaluate the performance of some supervised machine learning algorithms in spam detection. As determined by simulation results, five (5) machine learning algorithms were selected for this study, including the C4.5 decision tree algorithm, Naive Bayes, Support Vector Machine, Radial Basis Function Classifier, and K-nearest Neighbor. These supervised machine learning algorithms were chosen based on their strengths in predicting class in binary and multi-class problems. They are capable of dealing with high-dimensional data as well as small datasets. They have been certified in the literature to have a high level of accuracy in making real-time predictions as well as good generalization, even when the number of dimensions exceeds the number of samples [31]. A wrapper-based method was used for the selection and parameter determination of the selected models. To find the parameter values that drive each selected algorithm in closest agreement with the data, each feature subset of the chosen algorithms was evaluated on their performance quality using a greedy search strategy, i.e., Best First Search. Finally, it selected the optimal combination of features for each of the machine learning algorithms using the following pseudo code.

- *Begin*
- *Initialize all features*
- *Analyse a subset of features*
- *Apply Machine Language Algorithm*
- *Gauge model performance*
- *Repeat steps iii-v until optimal set of features is achieved*
- *End*

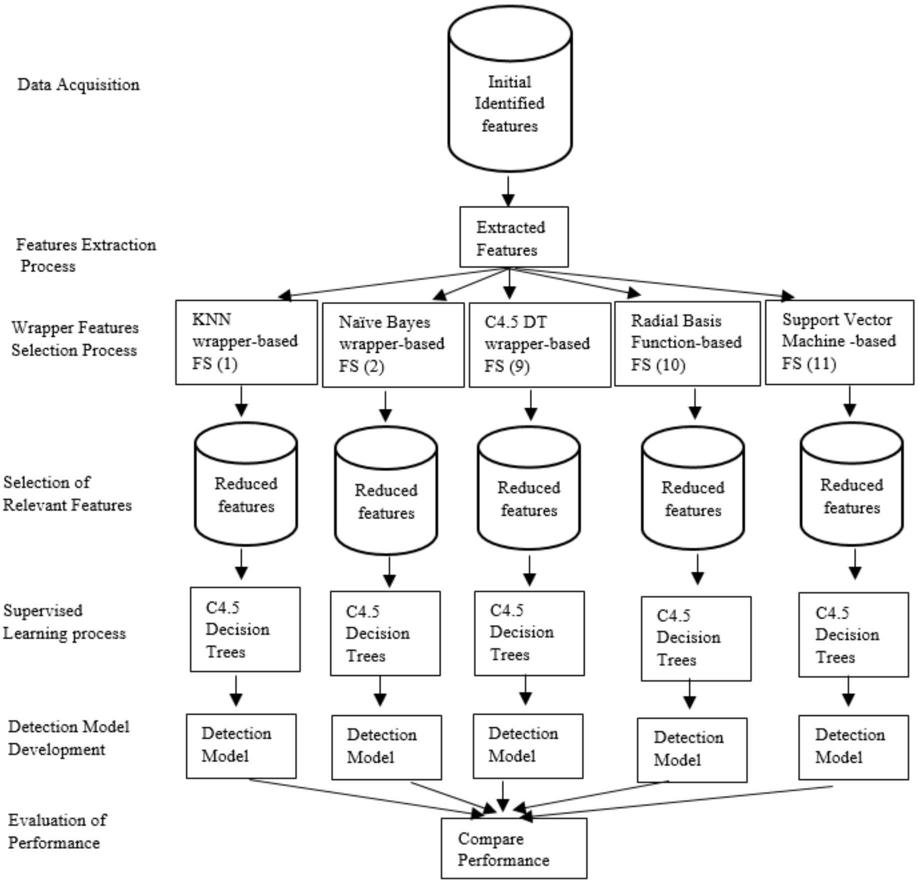

Data Acquisition
Initial Identified features
Extracted Features
Features Extraction Process
Wrapper Features Selection Process
KNN wrapper-based FS (1)
Naïve Bayes wrapper-based FS (2)
C4.5 DT wrapper-based FS (9)
Radial Basis Function-based FS (10)
Support Vector Machine -based FS (11)
Selection of Relevant Features
Reduced features
Reduced features
Reduced features
Reduced features
Reduced features
Supervised Learning process
C4.5 Decision Trees
C4.5 Decision Trees
C4.5 Decision Trees
C4.5 Decision Trees
C4.5 Decision Trees
Detection Model Development
Detection Model
Detection Model
Detection Model
Detection Model
Detection Model
Evaluation of Performance
Compare Performance

Fig.3. Conceptual view of the Robocall detection model

B. Filter- based Feature Selection Process

Three (3) filter-based feature selection methods, namely: correlation-based, information gain, and gain ratio-based, were employed in this study. Statistical techniques were used to evaluate the relationship between the input and the target variables. The evaluation scores were then used as the baseline to filter the required input variables in the detection model. The features were ranked from most relevant to least relevant for robocall detection. The most significant features were selected, and the rest were discarded.

3.2. *Mathematical Model Formulation*

In this study, mapping functions were used to express the process of model development. The training dataset *S*, which consists of the original features identified at the point of data identification and collection, is represented by X_i, where i, is the number of features existing in the original dataset of the audio recordings of human calls and robocalls, and X'_j consists of the features relevant for detecting robocalls where $j < i$. The mapping function, F in Equation 1 represents the feature selection process.

$$F : X_i \to X'_j \tag{1}$$

Where, X_i represents the original set of attributes collected and; X'_j represents the selected relevant features.

Following the process of feature selection, the new dataset belongs to X'_{jk} such that k is the number of host record cases collected in the original dataset. If n datasets are selected for training the model, the mapping in Equation (2) is used to formulate the model using the relevant variables:

$$\varphi : X_{jk} \to Y_k;\ \text{defined as: } \varphi\left(X_{jk}\right) = Y_k \text{ for all calls, } k \tag{2}$$

Where, X_{jk} the set of attributes, j for host records, k and Y_k is the target class (human calls or robocalls) of the record, k

Therefore, the supervised machine learning algorithm employed in this study is expected to determine the best fit for $\varphi \in \mathrm{H}$ (the set of all possible models) by minimizing the cost function defined in Equation (3).

$$\mathrm{L} : \mathrm{Y} \times \mathrm{Y} \to \mathrm{Z}\ ;\ \text{defined as: } \mathrm{L}\left(Y_a, Y_p\right) \tag{3}$$

Such that: Z is a set $[0,1]$ and Y_a, Y_p are the actual and predicted values of the target class, respectively.

Because the problem is a classification problem involving the identification of a value, values of 0 implied correct classification while values of 1 implied incorrect classification. As a result, the ability to correctly classify the detection of SPIT Robocalls is determined by the cost function defined in Equation (4).

$$\mathrm{L}\left(Y_a, Y_p\right) = \begin{cases} \text{correct classification} = 0 \\ \text{incorrect classification } = 1 \end{cases} \tag{4}$$

4. Results and Discussions

The proposed model was simulated and evaluated in the Waikato Environment for Knowledge Analysis (WEKA) environment. The detailed results are presented as follows:

4.1. *Dataset Analysis*

The acquired .*WAV* files were prepared for acoustic analysis by using the WarbleR R bundle's specan function. Specan measures 22 acoustic features on acoustic signals that have start and end times. The output from the preprocessed WAV files was saved as voice.csv, a 1.02 MB CSV file with 3168 rows and 21 columns (20 columns for each feature and one label column for the classification of robocalls or human calls). It was classified into 1586 robocalls and 1582 human calls. Following that, the mean value of all numeric attribute data was used as a threshold to convert all numeric values to their corresponding binary nominal values. For example, a variable's binary nominal values will be less than or equal to the variable's mean value k (< k) or larger than or equal to it (≥ k). Numeric variables were converted into nominal ones, which made them more manipulable than their numeric counterparts. As a result of time and CPU constraints, duration and peak frequency (peakf) were left out of the computation. There will be no difference in the duration (20s) and peak frequency of any of the records in this situation (0). Table 1 depicts a sample of the original dataset containing the extracted features, and Table 2 describes the various features extracted from the preprocessed dataset. Table 3 summarizes the numeric data types discovered among the variables in the voice dataset using the metrics: minimum, maximum, mean, and standard deviation for the dataset collected. A greedy search strategy was used for selecting attributes among the general set of attributes collected for each feature selection method chosen. The Greedy search strategy was used to select the optimal input features subset and to optimize the parameters of the models. Table 4 shows the relevant attributes and the optimal parameters for each model using the wrapper-based feature selection methods, while Table 5 shows the results of the filter-based feature selection methods as well as the search algorithm used and the relevant features identified for robocall detection.

Table 1. Sample of the Original Dataset Containing the Extracted Feature

	Meanfreq	sd	median	dfrange	modindx	……	Label
1	0.0597809849598081	0.0642412677031359	0.032026913372582	0	0	……	robot"
2	0.066008740387572	0.0673100287952527	0.040228734810579	0.0546875	0.046875	……	robot"
3	0.0773155026958227	0.0838294209445061	0.0367184586699814	0.0078125	0.0465116279069767	……	robot"
4	0.151228091724635	0.0721105872627985	0.158011187072716	0.5546875	0.247119078104994	……	robot"
5	0.135120387296677	0.0791461004935869	0.124656228727025	5.4765625	0.208273894436519	……	robot"
6	0.132786407306188	0.0795568659729794	0.119089848308051	2.71875	0.125159642401022	……	robot"
7	0.150762330177465	0.0744632052034133	0.160106382978723	5.3125	5.3046875	……	robot"
8	0.160514332166817	0.0767668848761485	0.144336775218427	0.53125	0.28393665158371	……	robot"
9	0.142239416807037	0.0780184624458836	0.138587443946188	2.15625	0.148272017837235	……	robot"
10	0.183667260279776	0.0406070122299704	0.182533632286996	3.4140625	0.166503214558134	……	human"
11	0.168794368168998	0.0858424332605394	0.188979980934223	5.875	0.268617021276596	……	human”
12	0.151770961732611	0.0891468392310416	0.185970149253731	0.192220052083333	1.5684211	……	human”
13	0.170655914426341	0.0812370506180178	0.184277108433735	0.7265625	0.336917562724014	……	human"
14	0.146023362615033	0.0925246277539745	0.183433734939759	2.984375	0.258924321751547	……	human”
15	0.131883734383811	0.0847340652461286	0.153707370737074	4.203125	0.161928712005248	……	human"
16	0.116220922114276	0.0892211424551605	0.0767580452920143	5.58421	1.08155	……	human"
17	0.142056255712406	0.0957984262823456	0.18373123659757	2.9296875	0.194758620689655	……	human"
18	0.143658744830027	0.090628260997323	0.184976167778837	3.5859375	0.311002178649237	……	human"
19	0.165508946001837	0.0928835369116316	0.183043922369765	0.546875	0.35	……	human"

Table 2. Description of Extracted Features

Variable	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Meanfreq	0.039	0.251	0.181	0.030
Sd	0.018	0.115	0.057	0.017
Median	0.011	0.261	0.186	0.036
Q25	0.000	0.247	0.140	0.049
Q75	0.043	0.273	0.225	0.024
IQR	0.015	0.252	0.084	0.043
Skew	0.142	34.725	3.140	4.241
Kurt	2.068	1309.613	36.568	134.929
sp.entb	0.739	0.982	0.895	0.045
Sfmmode	0.037	0.843	0.408	0.178
Centroid	0.000	0.280	0.165	0.077
Meanfun	0.039	0.251	0.181	0,030
Minfun	0.056	0.238	0.143	0.032
Maxfun	0.010	0.204	0.037	0.019
Meandom	0.103	0.279	0.259	0.030
Mindom	0.008	2.958	0.829	0.525
Maxdom	0.005	0.459	0.053	0.063
Dfrange	0.008	21.867	5.047	3.521
Modindx	0.000	21.44	4.995	3.520
Mode	0.000	0.932	0.174	0.119

Each wrapper-based feature selection algorithm selected a smaller number of relevant attributes from the total twenty (20) features extracted from the audio recording dataset for the dataset collected for this study. For example, Naive-Bayes' based feature selection selected two features, C4.5 Decision Trees-Based feature selection selected nine, SVM-Based feature selection selected eleven, RBF Classifier-Based feature selection selected ten, and KNN-Based feature selection selected only one. It was discovered that four (4) feature selection algorithms chose “IQR” and “minfun” as relevant features for this study. Three (3) feature selection methods identified “centroid” and “sfmmode” as relevant features for the work, while two (2) feature selection methods identified “meandom” and “skew” as relevant features for this study. SVM-based feature selection only identified “Q75” as a relevant feature, whereas the C4.5 decision tree feature selection algorithm identified “kurt” and “dfrange” as relevant features for the study (Table 3).

The filter-based feature selection methods, as opposed to the wrapper-based feature selection, ranked the features for robocall detection from most relevant to least relevant. For this study, it was discovered that the most relevant feature is the minimum fundamental frequency measured across the acoustic signal (minfun), and the least relevant feature is mode (mode frequency). The filter-based feature selection is only used to confirm the most important features in this study. They were not used in the simulation as feature selection methods (Table 4).

The feature selection process enabled us to create five distinct datasets based on the various wrapper-based feature selection algorithms used in this study.

Table 3. Numeric attribute data summarization

N	Name	Type	Description
1	duration	Numeric	length of signal
2	meanfreq	Numeric	meanfrequency (in kHz)
3	sd	Numeric	standard deviation of frequency
4	median	Numeric	medianfrequency (in kHz)
5	Q25	Numeric	first quantile (in kHz)
6	Q75	Numeric	third quantile (in kHz)
7	IQR	Numeric	interquantile range (in kHz)
8	skew	Numeric	skewness (see note in specprop description)
9	kurt	Numeric	kurtosis (see note in specprop description)
10	sp.entb	Numeric	spectral entropy
11	sfmmode	Numeric	spectral flatness
12	mode	Numeric	mode frequency
13	centroid	Numeric	frequencycentroid (seespecprop)
14	peakf	Numeric	peak frequency (frequency with highest energy)
15	meanfun	Numeric	average of fundamental frequency measured across acoustic signal
16	minfun	Numeric	minimum fundamental frequency measured across acoustic signal
17	maxfun	Numeric	maximum fundamental frequency measured across acoustic signal
18	meandom	Numeric	average of dominant frequency measured across acoustic signal
19	mindom	Numeric	minimum of dominant frequency measured across acoustic signal
20	maxdom	Numeric	maximum of dominant frequency measured across acoustic signal
21	dfrange	Numeric	range of dominant frequency measured across acoustic signal
22	modindx	Numeric	modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range

Table 4. Relevant attributes and optimal parameters identified using wrapper-based feature selection methods

Feature Selection Methods	Naïve Bayes- Based	C4.5 Decision Trees-Based	SVM-Based	RBF Classifier- Based	KNN-Based
Search Method	Best-First Search				
Variables Selected	IQR	IQR	IQR	IQR	Meanfreq
	Minfun	Q25	Q25	Minfun	
		Minfun	Minfun	sp.entb	
		sfmmode	sfmmode	sfmmode	
		centroid	centroid	centroid	
		meandom	meandom	Maxfun	
		mindom	Maxdom	Mindom	
		Kurt	skew	skew	
		Dfrange	Meanfreq	Modindx	
			Q75	Mode	
			Mode		

4.2. *Simulation Results*

The WEKA Explorer was used for the simulation. The training dataset, in its default file format (.arff), containing the relevant features, was subjected to ten runs of “ten-fold cross validation” using the five (5) supervised machine learning algorithms.

The confusion matrix was constructed for each classification model developed using the combination of "feature selection" and "supervised machine learning algorithms," showing the True Positive (TP), False Positive (FP), True

Negative (TN), and False Negative (FN) values over the ten (10) iterations, where;

- TP represents true positive and represents the situation in which a new call is a Robocall.
- TN stands for True Negative and represents a situation in which a call is not a Robocall but is classified as such by the model.
- FP stands for false positive and represents a benign call that the model classifies as a Robocall, and;
- FN stands for false negative and represents a robocall that the model classified as a benign call.

Each classification model's positive class was identified as robocalls, while the negative class was identified as human calls.

Table 5. Relevant attributes identified using filter-based feature selection methods

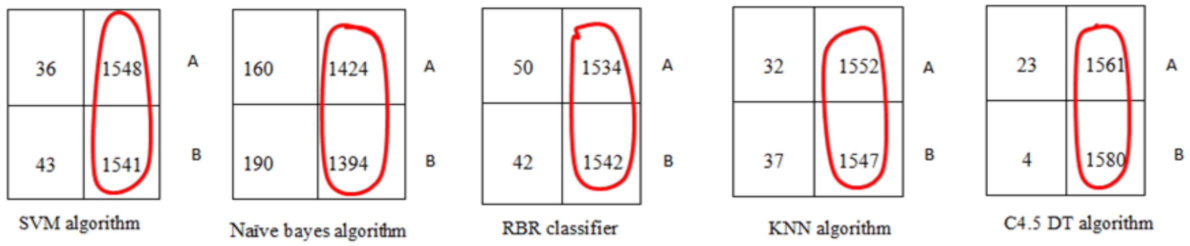
Feature Selection Methods	Correlation Based	Information Gain- Based	Gain ratio- Based
Search Method	Genetic Search	Ranker Search	Greedy Step-wise
Variables Selected	Minfun	Minfun	Minfun
	IQR	IQR	IQR
	Q25	Q25	Q25
	sp.entb	Sd	sp.entb
	Sd	sp.entb	Sd
	Sfmmode	Centroid	sfmmode
	Meanfreq	Sfmmode	Centroid
	Meanfun	Meanfun	Meanfun
	Median	Meanfreq	meanfreq
	Dfrange	Maxdom	Median
	Maxdom	Median	Maxdom
	Modindx	Skew	Skew
	Mindom	Dfrange	Kurt
	Centroid	Modindx	Mindom
	Meandom	Mindom	Dfrange
	Maxfun	Kurt	Modindx
	Kurt	Meandom	Maxfun
	Q75	Maxfun	meandom
	Skew	Q75	Q75
	Mode	mode	mode

A. Simulation Outcomes with all 20 Variables

The audio recording dataset, which included 20 variables (attributes) identified as relevant features for this study, was used as training data to create the first set of classifiers using the three supervised machine learning algorithms. A target class that defines the call's classification was labeled as robocalls and human calls for each call's record. The following are the results of the detection model developed using the various learning algorithms from a dataset containing 3168 call records (both robocalls and human calls):

- For the C4.5 decision tree algorithms, 1561 of 1584 are actual robocalls, and 1580 are correctly classified, for a total of 3141 correct classifications out of 3168 with a 99.15% accuracy.
- According to the Naive Bayes, 1424 of the 1584 calls are genuine robocalls, and 1394 of the 1584 calls are correctly classified, for a total of 2818 correct classifications out of 3168, with an accuracy of 88.95%.
- In the case of the Support Vector Machines, 1548 of the 1584 calls are actual robocalls, and 1541 of the 1584 calls are correctly classified, for a total of 3089 correct classifications out of 3168, with a 97.51% accuracy.
- For the K-Nearest Neighbors, 1552 of 1584 are actual robocalls and 1547 of 1584 are correctly classified, for a total of 3099 correct classifications out of 3168 with a 97.82% accuracy, and;
- For the Radial Basis Function Classifier, 1534 of 1584 are actual robocalls and 1542 of 1584 are correctly classified, for a total of 3077 correct classifications out of 3168 with a 97.1% accuracy.

Thus, using all 20 variables identified for this study, the C4.5 decision tree algorithm proved to be the most accurate (99.15%) detection model for robocall classification. Fig. 4 depicts the model formulation outcome when all 20 variables are used.



SVM algorithm

36	1548	A
43	1541	B

Naïve bayes algorithm

160	1424	A
190	1394	B

RBR classifier

50	1534	A
42	1542	B

KNN algorithm

32	1552	A
37	1547	B

C4.5 DT algorithm

23	1561	A
4	1580	B

A = Predicted as Robocalls

B = Predicted as Human calls

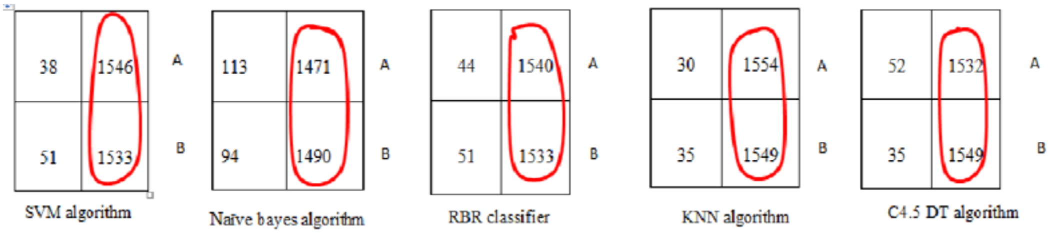
Fig.4. Simulation Results using all 20 variables

B. *Simulation Outcomes Using Selected Variables by C4.5 Decision Tree Algorithms.*

The dataset containing the nine (9) relevant variables was used to create the second set of classifiers. As a result, each learning algorithm created a classification model for voice calls based on the relevant attributes chosen in this case. The following are the results of the detection model developed using the various learning algorithms from a dataset containing 3168 call records (both robocalls and human calls):

- For the C4.5 decision tree algorithms, 1532 of 1584 calls are actual robocalls, and 1549 of 1584 calls are correctly classified, for a total of 3081 correct classifications out of 3168, with a 97.25% accuracy.
- For the Naive Bayes, 1471 of the 1584 calls are actual robocalls, and 1490 of the 1584 calls are correctly classified, for a total of 2961 correct classifications out of 3168, with a 93.47% accuracy.
- In the case of the Support Vector Machines, 1546 of the 1584 calls are actual robocalls, and 1533 of the 1584 calls are correctly classified, for a total of 3079 correct classifications out of 3168, with a 97.19% accuracy.
- For the K-Nearest Neighbors, 1554 of 1584 calls are actual robocalls, and 1549 of 1584 calls are correctly classified, for a total of 3103 correct classifications out of 3168, with a 97.95% accuracy, and;
- For the Radial Basis Function Classifier, 1540 of 1584 are genuine robocalls, and 1533 are correctly classified, for a total of 3073 correct classifications out of 3168 with a 97.00% accuracy.

Thus, using the 9 identified variables, the K-Nearest Neighbors algorithm proved to be the most accurate detection model for robocalls (97.95%). Fig. 5 depicts the outcome of model formulation with nine (9) variables.



SVM algorithm

38	1546	A
51	1533	B

Naïve bayes algorithm

113	1471	A
94	1490	B

RBR classifier

44	1540	A
51	1533	B

KNN algorithm

30	1554	A
35	1549	B

C4.5 DT algorithm

52	1532	A
35	1549	B

A = Predicted as Robocalls

B = Predicted as Human calls

Fig.5. Simulation Results using 9 variables

C. *Simulation Outcomes Using Selected Variables by the K-Nearest Neigbours Algorithm*

The third set of classifiers was developed by using a dataset with one (1) variable selected by the K-Nearest Neighbors algorithm. As a result, each learning algorithm developed a classification model for voice calls based on the relevant features identified in this case. The following are the results of the detection model developed using the various learning algorithms from a dataset containing 3168 call records (both robocalls and human calls):

- For the C4.5 decision tree algorithms, 1489 of the 1584 calls are actual robocalls, and 1049 of the 1584 calls are correctly classified, for a total of 2024 correct classifications out of 3168 with an accuracy of 63.89%.
- According to the Naive Bayes, 1471 of the 1584 calls are actual robocalls, and 1490 of the 1584 calls are correctly classified, for a total of 2032 correct classifications out of 3168, or 64.14% accuracy.
- In the case of the Support Vector Machines, 1004 of the 1584 calls are actual robocalls, and 1043 of the 1584

calls are correctly classified, for a total of 2047 correct classifications out of 3168, with a 64.61% accuracy.

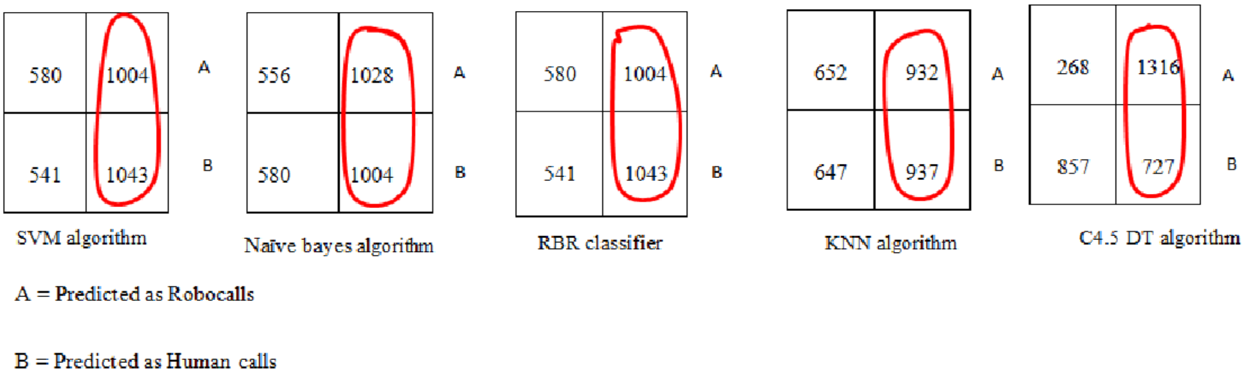
- For the K-Nearest Neighbors, 932 of 1584 are actual robocalls, and 937 of 1584 are correctly classified, for a total of 1869 correct classifications out of 3168 with a 58.99% accuracy, and;
- For the Radial Basis Function Classifier, 1316 of 1584 are actual robocalls, and 727 are correctly classified, for a total of 2043 correct classifications out of 3168 with a 64.49% accuracy.

Thus, using the identified variables, the Support Vector Machines algorithm proved to be the most accurate detection model for robocall classification (64.61%). Fig. 6 shows the results of the model formulation using only one (1) variable.

D. *Simulation Outcomes Using Selected Variables by the Naïve Bayes Algorithm*

The dataset containing the 2 variables selected by the Naive Bayes algorithm was used to develop the fourth set of classifiers. As a result, each learning algorithm developed a classification model for voice calls based on the relevant attributes chosen in this case. The detection model developed using the various learning algorithms produced the following results from a dataset containing 3168 call records (both robocalls and human calls).

- For the C4.5 decision tree algorithms, 1544 of the 1584 calls are actual robocalls, and 1523 of the 1584 calls are correctly classified, for a total of 3067 correct classifications out of 3168 with a 96.81% accuracy.
- For the Naive Bayes, 1543 out of 1584 are actual robocalls, and 1526 out of 1584 are correctly classified, for a total of 3069 correct classifications out of 3168 with a 96.88% accuracy.
- For the Support Vector Machines, 1550 of the 1584 calls are actual robocalls, and 1497 of the 1584 calls are correctly classified, for a total of 3047 correct classifications out of 3168 with a 96.18% accuracy.
- For the K-Nearest Neighbors, 1515 of 1584 are actual robocalls and 1520 of 1584 are correctly classified, for a total of 3035 correct classifications out of 3168 with an accuracy of 95.8%, and;
- For the Radial Basis Function Classifier, 1533 of 1584 are actual robocalls and 1533 of 1584 are correctly classified, for a total of 3066 correct classifications out of 3168 with an accuracy of 96.78%.



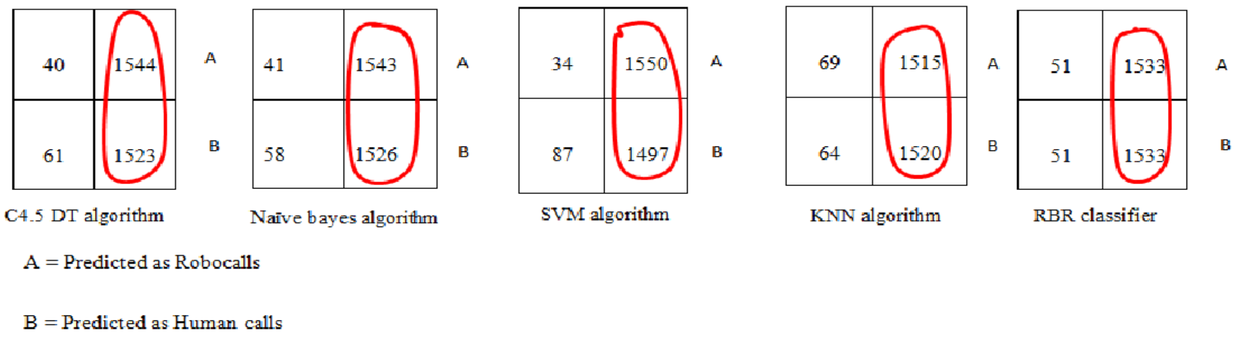
SVM algorithm			Naïve bayes algorithm			RBR classifier			KNN algorithm			C4.5 DT algorithm		
580	1004	A	556	1028	A	580	1004	A	652	932	A	268	1316	A
541	1043	B	580	1004	B	541	1043	B	647	937	B	857	727	B

A = Predicted as Robocalls

B = Predicted as Human calls

Fig.6. Simulation Results using 1 variable

Thus, using the two (2) identified variables, the Naïve Bayes algorithm proved to be the most accurate detection model for robocall classification (96.88%). Fig. 7 shows the results of the model formulation using two (2) variables.



C4.5 DT algorithm			Naïve bayes algorithm			SVM algorithm			KNN algorithm			RBR classifier		
40	1544	A	41	1543	A	34	1550	A	69	1515	A	51	1533	A
61	1523	B	58	1526	B	87	1497	B	64	1520	B	51	1533	B

A = Predicted as Robocalls

B = Predicted as Human calls

Fig.7. Simulation Results using 2 variables

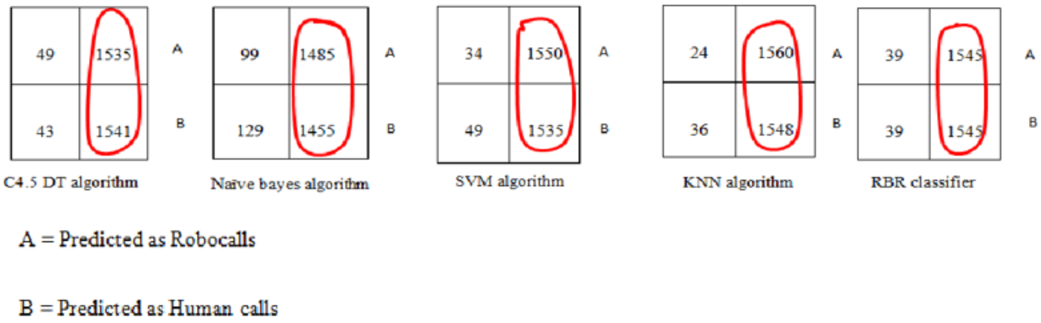
E. *Simulation Outcomes Using Selected Variables by the Radial Basis Function Algorithm*

The fifth classifier set was developed using the dataset containing the ten variables chosen by the Radial Basis

Function classifier. As a result, each learning algorithm created a classification model for voice calls based on the relevant attributes chosen in this case. The following are the results of the detection model developed using the various learning algorithms from a dataset containing 3168 call records (both robocalls and human calls):

- For the C4.5 decision tree algorithms, 1535 of 1584 calls are actual robocalls, and 1541 of 1584 calls are correctly classified, for a total of 3076 correct classifications out of 3168, with a 97.10% accuracy.
- For the Naive Bayes, 1485 of the 1584 calls are actual robocalls, and 1455 of the 1584 calls are correctly classified, for a total of 2940 correct classifications out of 3168 with a 92.80% accuracy.
- In the case of the Support Vector Machines, 1550 of the 1584 calls are actual robocalls, and 1535 of the 1584 calls are correctly classified, for a total of 3085 correct classifications out of 3168, with a 97.38% accuracy.
- For the K-Nearest Neighbors, 1560 of 1584 are actual robocalls and 1548 of 1584 are correctly classified, for a total of 3108 correct classifications out of 3168 with an accuracy of 98.11%; and
- The Radial Basis Function Classifier correctly classified 1545 out of 1584 actual robocalls, for a total of 3090 correct classifications out of 3168 with an accuracy of 97.54%.

Thus, using the ten (10) identified variables, the K-Nearest Neigbours algorithm proved to be the most accurate detection model for robocall classification (98.11%). Fig. 8 shows the results of the model formulation using ten (10) variables.



49	1535	A
43	1541	B

C4.5 DT algorithm

99	1485	A
129	1455	B

Naive bayes algorithm

34	1550	A
49	1535	B

SVM algorithm

24	1560	A
36	1548	B

KNN algorithm

39	1545	A
39	1545	B

RBR classifier

A = Predicted as Robocalls

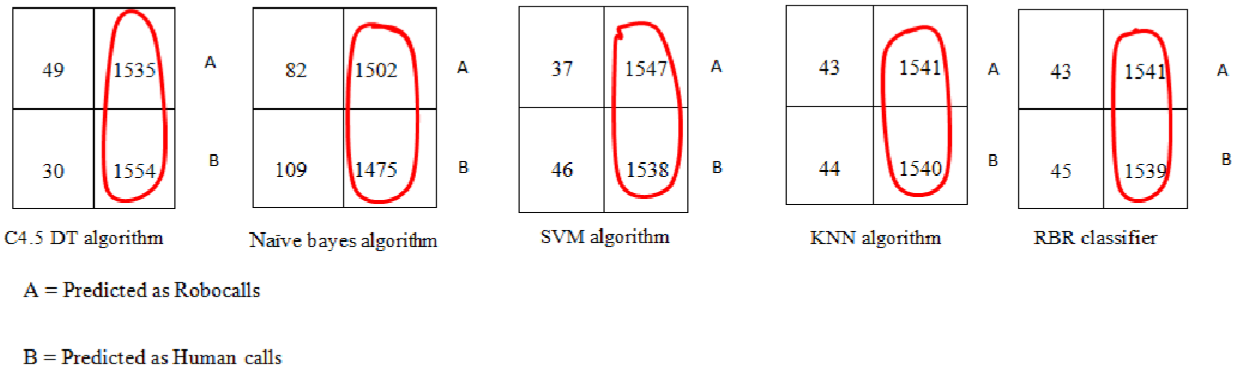
B = Predicted as Human calls

Fig.8. Simulation Results using 10 variables

F. Results of Model Formulation and Simulation Using Variables by Support Vector Machine based Feature Selection (11)

The dataset containing the variables (10) selected by the Support Vector Machine was used to develop the sixth set of classifiers. As a result, each learning algorithm developed a classification model for voice calls based on the relevant attributes chosen in this case. The following are the results of the detection model developed using the various learning algorithms from a dataset containing 3168 call records (both robocalls and human calls):

- For the C4.5 decision tree algorithms, 1535 of the 1584 calls are actual robocalls, and 1554 of the 1584 calls are correctly classified, for a total of 3089 correct classifications out of 3168 with a 97.51% accuracy.
- For the Naive Bayes, 1502 of the 1584 calls are actual robocalls, and 1475 of the 1584 calls are correctly classified, for a total of 2977 correct classifications out of 3168 with a 93.97% accuracy.
- For the Support Vector Machines, 1547 of the 1584 calls are actual robocalls, and 1538 of the 1584 calls are correctly classified, for a total of 3085 correct classifications out of 3168 with a 97.38% accuracy.



49	1535	A
30	1554	B

C4.5 DT algorithm

82	1502	A
109	1475	B

Naive bayes algorithm

37	1547	A
46	1538	B

SVM algorithm

43	1541	A
44	1540	B

KNN algorithm

43	1541	A
45	1539	B

RBR classifier

A = Predicted as Robocalls

B = Predicted as Human calls

Fig.9. Simulation Results using 11 variables

- For the K-Nearest Neighbors, 1541 of 1584 are actual robocalls and 1540 of 1584 are correctly classified, for a total of 3181 correct classifications out of 3168 with an accuracy of 97.25%, and;
- For the Radial Basis Function Classifier, 1541 of 1584 are actual robocalls and 1539 of 1584 are correctly classified, for a total of 3080 correct classifications out of 3168 with an accuracy of 97.22%.

Thus, using the eleven (11) identified variables, the C4.5 decision tree algorithm proved to be the most accurate detection model for robocall classification (97.51%). Fig. 9 shows the results of the model formulation using 11 variables.

4.3. Performance Evaluation Results

The performances of the developed models were evaluated using precision, recall, F-measure, accuracy, and Root Mean Square Error (RMSE) as follows:

- Accuracy denotes the correctness or efficiency of the classification, i.e., how much of the total number of cases was correctly classified by the classifiers.

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

- False Alarm Rate (FAR) measures the incorrectly classified negative cases. It is also known as the likelihood of false detection.

$$FAR = \frac{FP}{TN + FP} \tag{6}$$

- Precision measures the exact proportion of each class's correct classification, i.e., the proportion of voice calls that were predicted to be robocalls and were, in fact, robocalls. This is calculated as follows:

$$precision = \frac{TP}{TP + FP} \tag{7}$$

- Sensitivity measures the ability of the model to correctly classify the robocalls cases, i.e., the likelihood that a call will be a robocall. This is calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

Table 6 shows the detailed evaluation results. It was revealed that the C4.5 decision tree algorithm demonstrated a consistent and higher and thus performed better as the number of relevant features selected increased. For example, when using all 20 selected features, the C4.5 decision tree algorithm correctly classified 99.15% of the total dataset, with 97.72% of robocalls correctly classified and 97.29% of human calls correctly classified. Thus, the C4.5 decision tree algorithm outperformed the SVM, RBF, KNN, and NB algorithms when all 20 variables were used, as well as when 11 relevant variables were selected using an SVM-based feature selection method. According to Fig. 10, when the number of features is between 0 and 5, the C4.5 decision tree has an accuracy of less than 97%, and when the number of features is greater than 10, the C4.5 decision tree has an accuracy of greater than 97%.

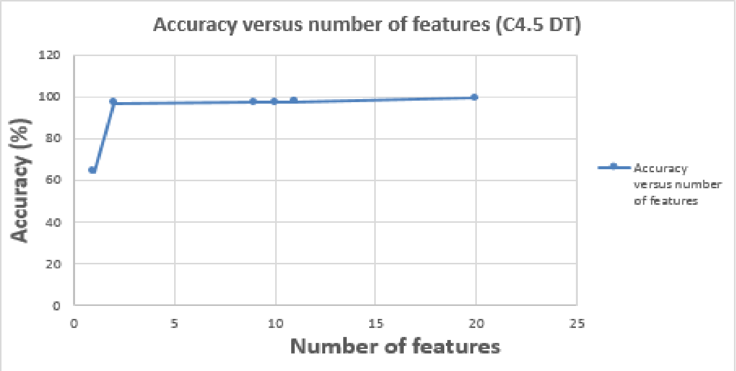
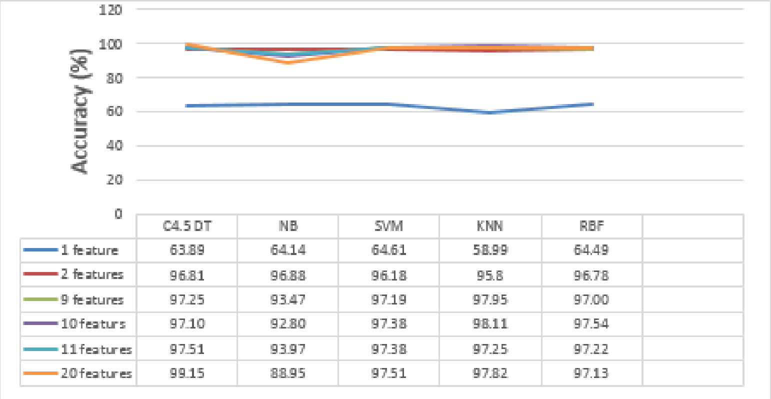


Fig.10. C4.5 DT Accuracy graph



	C4.5 DT	NB	SVM	KNN	RBF	
1 feature	63.89	64.14	64.61	58.99	64.49	
2 features	96.81	96.88	96.18	95.8	96.78	
9 features	97.25	93.47	97.19	97.95	97.00	
10 featurs	97.10	92.80	97.38	98.11	97.54	
11 features	97.51	93.97	97.38	97.25	97.22	
20 features	99.15	88.95	97.51	97.82	97.13	

Fig.11. Comparative study of the C4.5 decision tree with others algorithm

Table 6. Evaluation Results of Robocalls detection models

Feature Selection Method	Supervised Machine Learning Algorithm	Correct Classification	Accuracy (%)	Sensitivity	False Alarm Rate	Precision	Model Build Time (sec)
None (All 20 features)	C4.5 Decision Trees	3141	99.15	0.991	0.009	0.992	0.15
	Naïve Bayes	2818	88.95	0.890	0.110	0.890	0.03
	SVM	3089	97.51	0.975	0.025	0.975	0.48
	KNN	3099	97.82	0.978	0.022	0.978	0.00
	RBF	3077	97.13	0.971	0.029	0.971	1.68
C4.5 DT Wrapper Based FS (9)	C4.5 Decision Trees	3081	97.25	0.973	0.027	0.973	0.10
	Naïve Bayes	2961	93.47	0.935	0.065	0.935	0.02
	SVM	3079	97.19	0.972	0.028	0.972	0.35
	KNN	3103	97.95	0.979	0.021	0.979	0.00
	RBF	3073	97.00	0.970	0.030	0.970	0.80
KNN Wrapper Based FS (1)	C4.5 Decision Trees	2024	63.89	0.639	0.361	0.718	0.07
	Naïve Bayes	2032	64.14	0.641	0.359	0.641	0.01
	SVM	2047	64.61	0.646	0.354	0.646	0.14
	KNN	1869	58.99	0.588	0.412	0.588	0.00
	RBF	2043	64.49	0.645	0.355	0.668	0.67
Naïve Bayes Wrapper Based FS (2)	C4.5 Decision Trees	3067	96.81	0.968	0.032	0.968	0.03
	Naïve Bayes	3069	96.88	0.969	0.031	0.969	0.00
	SVM	3047	96.18	0.962	0.038	0.962	0.08
	KNN	3035	95.8	0.958	0.042	0.958	0.00
	RBF	3066	96.78	0.968	0.032	0.968	0.33
Radial Basis Function Wrapper Based FS (10)	C4.5 Decision Trees	3076	97.10	0.971	0.029	0.971	0.10
	Naïve Bayes	2940	92.80	0.928	0.072	0.928	0.01
	SVM	3085	97.38	0.974	0.026	0.974	0.20
	KNN	3108	98.11	0.981	0.019	0.981	0.00
	RBF	3090	97.54	0.975	0.025	0.975	1.58
Support Vector Machine Wrapper Based FS (11)	C4.5 Decision Trees	3089	97.51	0.975	0.025	0.975	0.08
	Naïve Bayes	2977	93.97	0.940	0.060	0.940	0.02
	SVM	3085	97.38	0.974	0.026	0.974	0.23
	KNN	3081	97.25	0.973	0.027	0.973	0.00
	RBF	3080	97.22	0.972	0.028	0.972	0.66

The conclusion is that the greater the number of relevant features, the greater the accuracy. Fig. 11 compares the accuracy of the C4.5 decision tree to the other classifiers used in the evaluation. It is demonstrated that the SVM performed better in the first case (64.6%), the NB performed better in the second case (96.88%), the KNN performed better in the third and fourth cases (97.95% and 98.11%), and the C4.5 decision tree performed better in the last two cases (97.51%) and 99.15%). In Table 5, the C4.5 decision tree outperforms the first and second cases in terms of precision (0.718) and (0.968), with costs of 0.07 and 0.03 seconds, respectively. KNN performed better in the third and fourth cases (0.979) and (0.981), with costs of 0.00 and 0.00 seconds, respectively. The C4.5 decision tree performed better (0.975) and (0.992) with lower costs (0.08) and 0.15) in the fifth and sixth cases, respectively.

5. Conclusions

Robocalls are a type of spam over Internet telephony that has become extremely difficult for researchers to detect. This study used audio recordings to develop a detection model that can distinguish between robocalls and human calls. From the feature selection process, interquantile range (IQR) and minimum fundamental frequency (minfun) were determined to be the optimal parameters for robocall detection using C4.5 decision trees. Additionally, the first quantile (Q25) as well as the spectral flatness (sfmmode) have been found to be highly significant factors as well. However, the C4.5 decision tree algorithm identified mode frequency (mode) as being the least relevant variable to robocall detection.

According to the number of variables used in the formulation of the detection model, the detection model developed using the dataset showed good results. Identification of the variables by means of a detection model as well as construction of a decision tree using the variables can provide insight into relationships between the variables as they relate to classification models developed using the 20 variables as well as 11 variables, respectively.

The performance evaluation results showed that the higher the number of relevant features selected, the better the C4.5 decision tree results. The C4.5 decision tree algorithm outperformed the Support Vector Machine, Radial Basis Function classifier, K-nearest Neighbour, and Naïve Bayes algorithms while using 20 and 11 relevant variables. The C4.5 decision tree algorithm showed consistent and higher performance when using all 20 features extracted from the original dataset for this study; 99.15% of the total dataset was correctly classified, with 97.72% of robocalls correctly classified and 97.29% of human calls correctly classified, owing to a true positive rate, false alarm rate, and precision of 0.991, 0.009, and 0.992, respectively.

The model's ability to correctly detect and classify voice calls as either robocalls or human calls was up to 99.15%, thus showing that the C4.5 decision tree machine learning approach could be used to detect robocalls effectively. The limitation of the model is that it takes more time for testing and training, and a little change in the data leads to a serious change in the structure of the model. Future work will focus on detecting robocalls using an ensemble approach to machine learning.

Acknowledgment

This research was funded by the TETFund Research Fund and the Africa Centre of Excellence, OAK-Park

References

[1] I.T. Javed, K. Toumi, F. Alharbi, T. Margaria, and N.Crespi, “Detecting Nuisance Calls over Internet Telephony Using Caller Reputation. *Electronics,* 10 (3), pp.353, 2021, http://doi.org/10.3390/electronics10030353

[2] M. Snider “Robocalls rang up a new high in 2019, two or more daily is average in some states”. Available at https://www.usatoday.com/story/tech/2020/01/15/robocalls-americans-got-58-5-billion-2019/4476018002, 2020.

[3] H. Tu, A. Doupe, Z. Zhao, and G.-J., Ahn “Sok: Everyone hates robocalls: A survey of techniques against telephone spam,” *In the proceedings of the 2016 IEEE Symposium on Security and Privacy (SP),* 9781509008247, pp. 320–338, 2016, http://dx.doi.org/10.1109/SP.2016.27

[4] H. Jarral, F. Mehmood, and A, Ali “Centralized Spam over Internet Telephony (SPIT) control on VoIP”. *International Journal of Scientific and Research Publications*, 7(2), pp 118-121, 2017.

[5] J. Pandit, R.P. Liu, and M. Ahamad, "Applying Deep Learning to Combat Mass Robocalls," *in the proceeding of the 2021 IEEE Security and Privacy Workshops (SPW),* 978166543732, pp. 63-70, 2021, http://dx.doi.org/10.1109/SPW53761.2021.00018

[6] T. Wadhwa. Why robocalls are about to get more dangerous. CNN Business, Available at: *https://edition.cnn.com/2018/10/16/perspectives/robocalls-voice-manipulation-tech/index.html,* 2018.

[7] I. Sherman, J. Bowers, K. McNamara Jr, J. Gilbert, J. Ruiz, and P. Traynor. “Are You Going to Answer That? Measuring User Responses to Anti-Robocall Application Indicators”. *In the Proceedings of the 2020 Network and Distributed System Security Symposium,* 1891562614, 2020, http://dx.doi.org/10.14722/ndss.2020.24286

[8] L. Junda, K. Naveen, and L. Shi, "Robocall and fake caller-id detection", Technical Disclosure Commons, (December 01, 2017) https://www.tdcommons.org/dpubs_series/845, 2017.

[9] I. N. Sherman, J. D. Bowers, L. -L. Laborde, J. E. Gilbert, J. Ruiz and P. G. Traynor, "Truly Visual Caller ID? An Analysis of Anti-Robocall Applications and their Accessibility to Visually Impaired Users," *2020 IEEE International Symposium on Technology and Society (ISTAS*), 9781665415071, pp. 266-279, , 2020, http://dx.doi.org/10.1109/ISTAS50296.2020.9462185

[10] J. Xing, M. Yu, S. Wang, Y. Zhang, and Y. Ding, “Automated Fraudulent Phone Call Recognition through Deep Learning”. *Wireless Communications and Mobile Computing,* 2020, http://dx.doi.org/10.1155/2020/8853468

[11] B.O. Akinyemi, A.O. Amoo and E.A. Olajubu (2014), “An Adaptive Decision-Support Model for Data Communication Network Security Risk Management”. *International Journal of Computer Applications,* 106(8), pp.1-7, http://dx.doi.org/10.5120/18537-9752

[12] O.H. Odukoya, B.O. Akinyemi, M. Fofana, and G.A. Aderounmu, “Performance evaluation of user-behaviour techniques of web spam detection models. *Network and Complex Systems (NCS*), vol. 10, pp. 59-73, 2019,

http://dx.doi.org/10.7176/NCS/10-07

[13] Y. Zhang, H. Wu, J. Zhang, J. Wang and X. Zou, "TW-FCM: An Improved Fuzzy-C-Means Algorithm for SPIT Detection," *in proceedings of the 27th International Conference on Computer Communication and Networks (ICCCN)*, 9781538651568, pp. 1-9, 2018, http://dx.doi.org/10.1109/ICCCN.2018.8487369

[14] S. Prasad, E. Bouma-Sims, A. K. Mylappan, and B. Reaves, "Who's Calling? Characterizing Robocalls through Audio and Metadata Analysis." *In the Proceedings of the 29th USENIX Security Symposium.* August 12–14, pp. 397-414, 2020.

[15] K. B. Kealy, and P. I. Rosencrantz, "Arrangement for managing voice over IP (VoIP) telephone calls, especially unsolicited or unwanted calls". U.S. Patent No. 7,912,192. Washington, DC: U.S. Patent and Trademark Office, 2011.

[16] G. Vennila, M.S.K Manikandan., and M.N. Suresh, "Detection and prevention of spam over Internet telephony in Voice over Internet Protocol networks using Markov chain with incremental SVM". *International Journal of Communication Systems,* 30(11), pp. e3255, 2017, http://dx.doi.org/10.1002/dac.3255

[17] K. Rieck, S. Wahl, P. Laskov, P. Domschitz, and K.R. Müller, "A self-learning system for detection of anomalous sip messages". In: Schulzrinne H., State R., Niccolini S. (eds), principles, systems and applications of IP telecommunications, services and security for next generation networks, *Lecture Notes in Computer Science*, 5310, 2008, http://dx.doi.org/10.1007/978-3-540-89054-6_5

[18] D. Hoffstadt, E. Rathgeb, M. Liebig, R. Meister, Y. Rebahi and T. Q Thanh, "A comprehensive framework for detecting and preventing VoIP fraud and misuse," in proceedings of the 2014 *International Conference on Computing, Networking and Communications (ICNC)*, 9781479923588, pp. 807-813, 2014, http://dx.doi.org/10.1109/ICCNC.2014.6785441

[19] G. Vennila, and M. S. K. Manikandan "Detection of Human and Computer Voice Spammers Using Hidden Markov Model in Voice Over Internet Protocol Network". *Procedia computer science,* 115, pp. 588-595, 2017, http://dx.doi.org/10.1016/j.procs.2017.09.169

[20] R. J. B Chikha, T. Abbes, W. B. Chikha, and A. Bouhoula, "Behavior-based approach to detect spam over IP telephony attacks". *International Journal of Information Security*, 15(2), pp. 131-143, 2016, http://dx.doi.org/10.1007/s10207-015-0281-1

[21] A. Natarajan, A. Kannan, V. Belagali, V. N. Pai, R. Shettar, and P. Ghuli, "Spam Detection Over Call Transcript Using Deep Learning," in *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 2* (K. Arai, ed.), (Cham), pp. 138–150, Springer International Publishing, 2022, http://dx.doi.org/10.1007/978-3-030-89880-9_10

[22] M. Ghosh and P. Prabu, "Empirical analysis of ensemble methods for the classification of robocalls in telecommunications" *International Journal of Electrical and Computer Engineering (IJECE),* 9(4), pp. 3108~3114, 2019, http://doi.org/10.11591/ijece.v9i4.pp3108-3114

[23] S. Harun, T. H. Bhuiyan, S. Zhang, H. Medal and L. Bian, "Bot Classification for Real-Life Highly Class-Imbalanced Dataset," *in proceedings of the 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress*, pp. 565-572, 2017, http://dx.doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.102

[24] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro. "Hello? Who Am I Talking to?" A Shallow CNN Approach for Human vs. Bot Speech Classification". *In the proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 9781479981311, 2019, http://dx.doi.org/10.1109/ICASSP.2019.8682743.

[25] B. Elizalde and D. Emmanouilidou, "Detection of Robocall and Spam Calls using Acoustic Features of Incoming Voicemails" *in the Proceedings of Meetings of the 181st Meeting of the Acoustical Society of America,* 29 November - 3 December 2021, Vol. 45, 060004 (2022) Seattle, Washington, http://dx.doi.org/10.1121/2.0001533

[26] M. G Hibbard, "Hanging up too early: remedies to reduce robocalls". *Journal of Law, Technology & the Internet*, 5, pp. 79-112, 2014.

[27] M. Mahoney, "Dialing Back: How Phone Companies Can End Unwanted Robocalls", Technical report, Consumer Unions, Policy and action from Consumers reports. Available at: https://www.fcc.gov/consumers/guides/stop-unwanted-robocalls-and-texts , 2015.

[28] F. Staff. "Protecting consumer privacy in an era of rapid change–a proposed framework for businesses and policymakers". *Journal of Privacy and Confidentiality*, 3(1), pp. 67-140, 2011, http://dx.doi.org/10.29012/jpc.v3i1.596

[29] H. Li, X. Xu, C. Liu, T. Ren, K. Wu, X. Cao, ... and D. Song, "A Machine Learning Approach to Prevent Malicious Calls over Telephony Networks," *2018 IEEE Symposium on Security and Privacy (SP*), pp. 53-69, 2018, http://dx.doi.org/10.1109/SP.2018.00034

[30] S. M Gowri, G. S. Ramana, M. S. Ranjani and T. Tharani, "Detection of Telephony Spam and Scams using Recurrent Neural Network (RNN) Algorithm," *in proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS),* 9781665405201, pp. 1284-1288, 2021, http://dx.doi.org/10.1109/ICACCS51430.2021.9441982

[31] Reena Sharma, Gurjot Kaur,"E-Mail Spam Detection Using SVM and RBF", International Journal of Modern Education and Computer Science, Vol.8, No.4, pp.57-63, 2016.

Authors' Profiles

Bodunde O. Akinyemi holds B.Tech (2005) in Computer Science from Ladoke Akintola University Ogbomosho, M.Sc. (2011) and Ph.D. (2014) in Computer Science from Obafemi Awolowo University Ile-Ife. She is a member of the International Association Engineers, IEEE, Nigeria Computer Society (NCS) and Computer Professional Registration Council of Nigeria (CPN). She is a Senior Lecturer and member of the Data Communication Group in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. Her major research areas are in Data Communication, Network Security and Performance management, Software Development and BlockChain Technology.

Helen O. Odukoya is a Lecturer I at Obafemi Awolowo University, Ile –Ife. Her research interests are Cybersecurity, Data mining and Data Communication and Networking (DCN).

Mistura L. Sanni obtained her BSc, M.Sc. (Computer Engineering) in 1992 and 2006 respectively from Obafemi Awolowo University, Ile-Ife. from and Ph.D. (Computer Engineering) from the International Islamic University, Malaysia in 2015 and currently a Senior Lecturer in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. She specializes in hardware design and data communication Networks. She is a certified COREN, NCS, CPN member.

Gilbert Sewagnon is a postgraduate student of Obafemi Awolowo University, Ile-Ife, Nigeria. His current research interest includes cyber security and Data mining.

Ganiyu A. Aderounmu is a professor of Computer Science and Engineering from Obafemi Awolowo University, Ile-Ife, Nigeria. He is a Full member of the Nigeria Society of Engineers (NSE) and also a registered Computer Engineer with Council for Regulation of Engineering Practice in Nigeria (COREN). He is also a Full member of Nigeria Computer Society (NCS) and Computer Professional Registration Council of Nigeria (CPN). He has over 30years of experience in teaching and research. He is an author of many journal articles in Nigeria and abroad. His special interest includes computer communication and network. He is a visiting Research Fellow to the University of Zululand, Republic of South Africa. He was the former head of the Department of computer Science & Engineering, former Dean, Faculty of Technology, former President of Nigeria Computer Society (NCS), and the former Director of Information Technology and Communication Unit (INTECU).

How to cite this paper: Bodunde O. Akinyemi, Oluwatoyin H. Odukoya, Mistura L. Sanni, Gilbert Sewagnon, Ganiyu A. Aderounmu, "Performance Evaluation of Machine Learning-Based Robocalls Detection Models in Telephony Networks", International Journal of Computer Network and Information Security(IJCNIS), Vol.14, No.6, pp.37-53, 2022. DOI:10.5815/ijcnis.2022.06.04