International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

A New Hybrid Strategy for Malware Detection Classification with Multiple Feature Selection Methods and Ensemble Learning Methods

P. HarshaLatha, R. Mohanasundaram

Abstract: ***A dramatic increase in malware in our day-to-day life causes a noteworthy problem in cyber security. The traditional approaches and signature-based models are not sufficient to defense with the new malware. To achieve zero-day attacks of malware, these approaches are not much competent to face new malware. To enhance the compete for the mechanism of classifying new malware the machine learning approaches are highly effective. To classify new malware with the high dimensionality of data leads to reduce the quality of output and low-performance results. In this paper, we propose a new hybrid strategy that combines the power of feature selection methods along with ensemble learning methods to improve accuracy for high dimensionality of data. This hybrid approach having three stages, preprocessing, feature selection and classification. Three different types of feature selection methods: ExtraTreesClassifier, Percentile and KBest feature selection methods are used to select the best features (dimensionality reduction) and four ensemble classifiers: AdaBoost, Gradient Boosting, Random Forest and Bagging are used for classification. The accuracy of ensemble classifiers are increased with this hybrid model and produces better results of classification with 91.50% accuracy. For dealing with the high dimensionality of data this hybrid approach is very effective and gives better results.***

Keywords: Hybrid Model, Dimensionality Reduction, Machine Learning, Feature Selection, Classification, Malware detection, Ensemble Learning.

I. INTRODUCTION

According to a recent survey of malware statistics [1] by comparitech, there is a drastic growth of malware and becomes more harmful to the internet and arises major concerns in security. Malware effects more on the internet especially in websites and there are millions of websites are blocked due to FormJacking attacks. To achieve zero-day attacks in recent decades become more crucial. For identifying the new malware there is a need for classification. The traditional approaches of classification [2] are not overcome some issues in classification because of advancements in malware growth. Machine learning approaches bring more advanced techniques in malware detection classification [3]. To improve classification accuracy several approaches are newly implemented in recent decades. The feature selection methods [4] support at most in malware detection classification and it provides the more advantages like it is a solution to the over-fitting problem, it is best to fit for dimensionality reduction, it reduces the processing time by removing irrelevant or unwanted features from the dataset. The other enhancement of machine learning is ensemble learning. The overview of ensemble learning methods is discussed in Section 3.

In this paper, we use more advanced techniques of feature selection methods and machine learning methods (Ensemble learning) and proposed a hybrid framework to improve classification accuracy. In Section 2, existing literature works are discussed. The overview of feature selection methods and ensemble learning methods are explained in Section 3. In Section 4, the new hybrid framework of malware detection classification is explained in detail. The experimental results, comparison and discussion carried out in Section 5. The conclusion of this paper is defined in Section 6.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
P. HarshaLatha*, Research Scholar, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India. E-mail: harsha17latha@gmail.com
R. Mohanasundaram, Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu, India. E-mail: mohanasundaramr@vit.ac.in

II. RELATED WORKS

Khammas et. al [5] presents a work of comparative study with different feature selection methods and machine learning methods. This work gets better results with the combination of Principal Components Analysis (PCA) and Support Vector Machine (SVM) classifier about 97% accuracy.

Yerima et. al [6] presents work to improve the accuracy rate of malware detection classification from 97-99%. [7] presents a work with five feature selection methods and five machine learning classifier Naïve Bayes (NB), Multi- Layer Perceptron (MLP), K-nearest Neighbour (KNN), Random Forest (RF) and Decision Tree(J48). The dataset is from the previous existing works. The overall best accuracy is 83% achieved with MLP classifier.

Narudin et. al [8] defines the work with the Genome dataset in the WEKA tool. The feature selection methods and five machine learning classifiers KNN, MLP, J48, RF and Bayes Network are used for classification. They got the best results with 99.9% accuracy with the RF classifier.

[9] presents a work with three different datasets, feature selection and ensemble classifier SVM-AR is used for classification. The better accuracy is obtained with ensemble classifier SVM-AR (Association Rules) and less execution time.

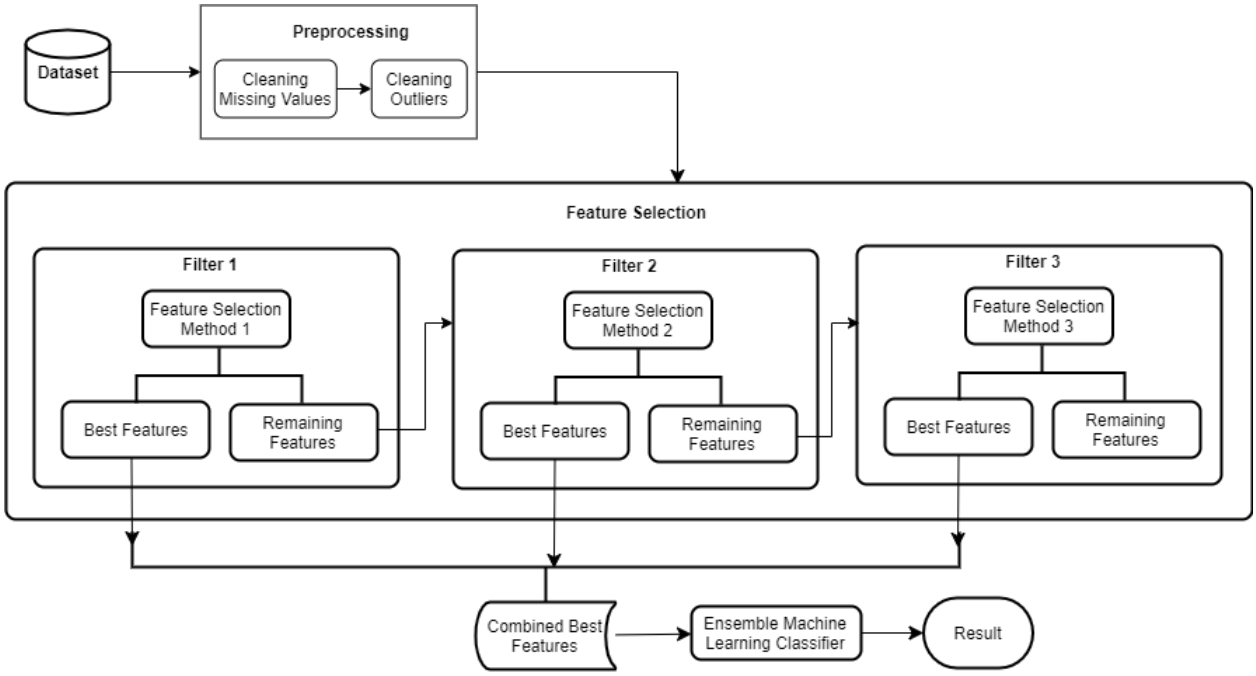

Fig. 1.Proposed Model for Hybrid Approach

III. OVERVIEW OF METHODS

The Feature Selection methods and Ensemble learning methods that are used in the hybrid framework of this paper are explained below.

A. Feature Selection Methods

Feature selection methods process the features to remove irrelevant features and it forms the best feature set that leads to improving high accuracy of classification [10-12]. The feature selection methods used in this paper are ExtraTreesClassifier, Percentile and KBest methods which are defined below.

ExtraTreesClassifier is an ensemble classifier used for feature selection. It is constructed using the training sample. It is a similar kind of Random Forest ensemble machine learning classifier[13]. The best features [14] are selected based on the formulae of Information Gain and Entropy which are (1) and (2) given below. It aggregates the several correlated decision trees to "forest".

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{S} Entropy(Sv) \quad (1)$$

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2(pi) \quad (2)$$

Where,

S: a training set

p_i : the proportion of rows with output label is i

c: Number of Unique Class labels

Percentile Method [15] is a feature selection method used to select the best features basing on the percentiles scores of features.

KBest [16] is a feature selection method used to select the best features basing on the k value of the highest scores.

B. Ensemble Learning

Ensemble learning is a collection of basic models of machine learning. It aggregates the solution from the decisions of different sample sets of basic models and gives the optimization solution to the problem [17, 18]. The ensemble-based machine learning classifiers used in this hybrid model is AdaBoost, Random Forest (RF) [19], Gradient Boosting, and Bagging [20].

IV. PROPOSED WORK

The proposed model contains three different stages. First, the preprocessing of the dataset, which means cleaning or removing missing values and outliers present in the dataset. Without preprocessing the dataset gives inaccurate data which leads to inaccurate results. Preprocessing is done WEKA tool, for finding any missing values or outliers. After preprocessing, now the dataset is ready for the feature selection process. The second process of this proposed model is feature selection; here feature selection methods are applied in three different filters. The third process contains the ensemble learning classifiers for classification. The Pseudo-code for the proposed model is given below.

1. Select Dataset A
2. Preprocess dataset A and removes missing values and outliers from the dataset A
3. Impute feature selection method on dataset A
4. Keep best feature subset FS1 from dataset A and remaining feature subset RFS separately
5. Impute feature selection method on RFS
6. Repeat the process from step 3 to step 6 for all three feature selection methods
7. Collect FS2, FS3

8. Combines all the feature subsets FS1, FS2, and FS3 to FS that to unique features, FS <-{FS1, FS2, FS3}
9. Build an Ensemble Machine Learning classifier model on FS.
10. Compute and Compare results.

The overview of proposed hybrid model is shown in Fig. 1and it is implemented in python script.

Dataset Description

The dataset for this hybrid approach is collected from the Internet. The dataset contains large dimensionality with 85 features and 3610 rows. Among them 693 files are Benign and the remaining files are malware files.

Feature Selection

Feature Selection is the process of selecting the best features among total features for classification which is most relevant. The feature selection process is implemented in Python. This hybrid proposed model contains a total of three filters. In each filter, the most relevant features are filtered and the remaining features are forwarded to the other frequent filter.
In Filter 1, the preprocessed dataset is considered for feature selection. The most relevant features are selected using the feature selection method ExtraTreesClassifier. After applying the ExtraTreesClassifier feature selection method on the preprocessed dataset and it produces output as 25 best features among total 85 features with feature importance values as scores defined in Fig. 3 and the overview of filter 1 is defined in Fig. 2. These kept features are kept in to aside for further processing, and the remaining 72 features are forwarded to the next filter i.e., second filter.

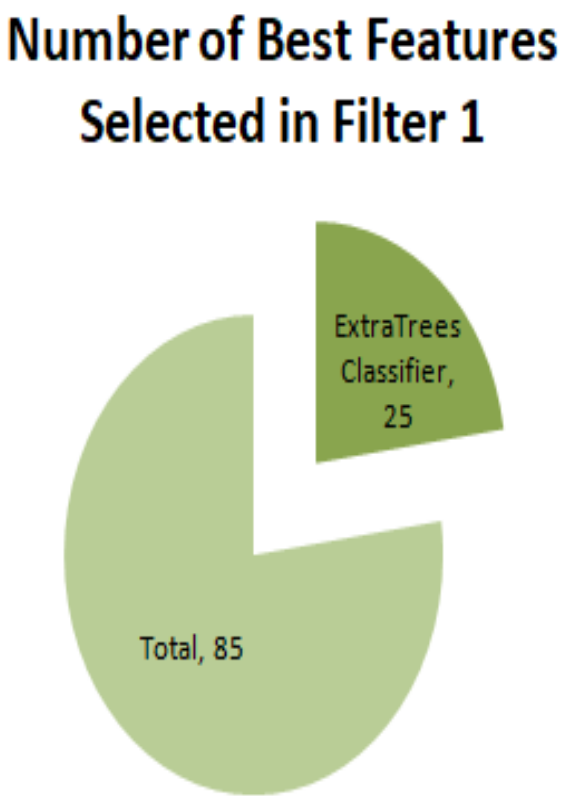

Fig. 2.Feature Selection using ExtraTrees Classifier

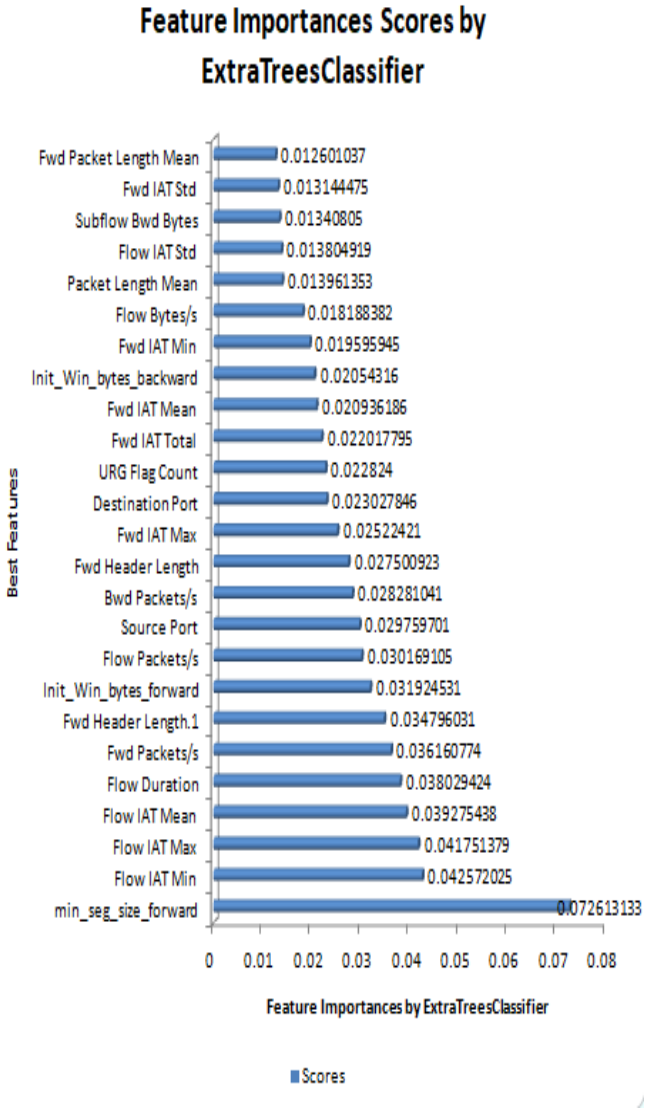

Fig. 3.Feature Importances of Best Selected Features by ExtraTreesClassifier

In Filter 2, the remaining features of Filter 1 are processed using feature selection method Percentile. The percentile method selects 6 best and relevant features from 60 features and the remaining 54 features are forwarded to the third Filter. The overview of feature selection by the percentile method is shown in Fig. 4. The feature importance scores of the 6 best features selected by the percentile method are shown in Fig. 5.
In Filter 3, the remaining features of Filter 2 are processed by using the KBest feature selection method and it selects 5 best features from 54 features. The overview of feature selection in Filter 3 is shown in Fig. 6 and the feature importances of these 5 best features are shown in Fig. 7.
The overview of the overall process of feature selection by the three filters is shown in Fig. 8.
After the feature selection process, the best features selected by ExtraTreesClassifier with 25 features, 6 features by Percentile and 5 features by KBest, all these features are combined and forms a dataset FS as mentioned in the pseudo-code, for further process of modeling.

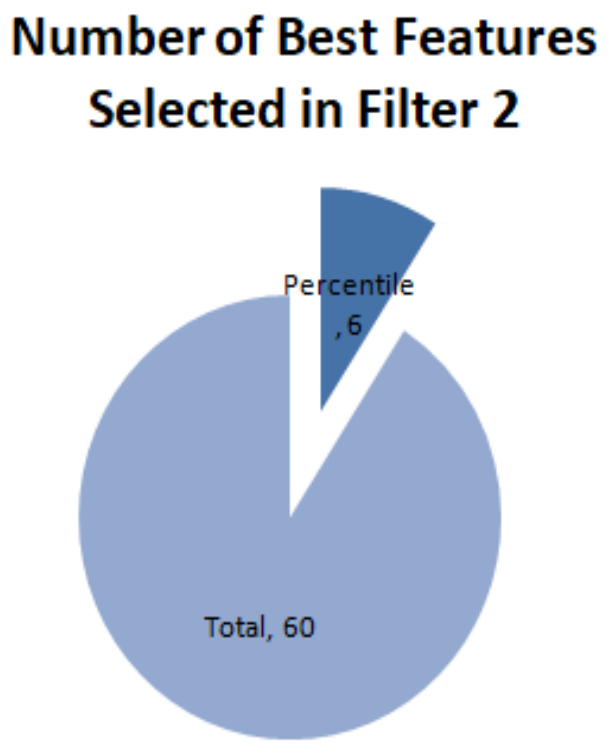

Fig. 4.Overview of Best Selected Features by Percentile Method

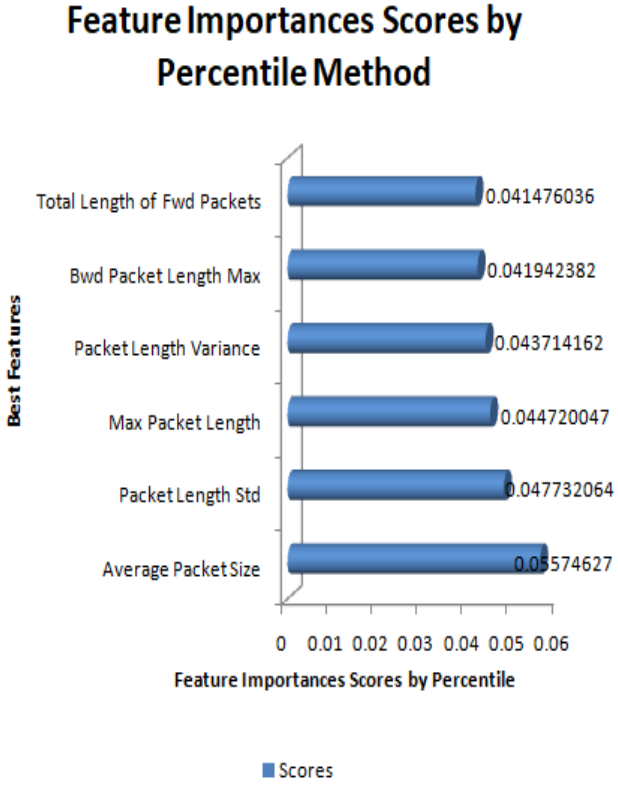

Fig. 5.Feature Importances of Best Selected Features by Percentile Method

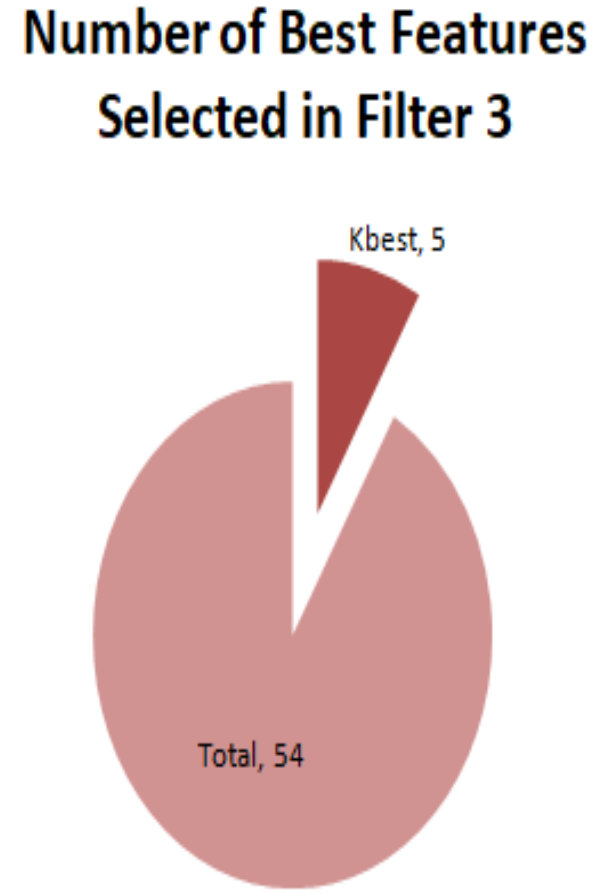

Fig. 6.Overview of best feature selection by KBest Method

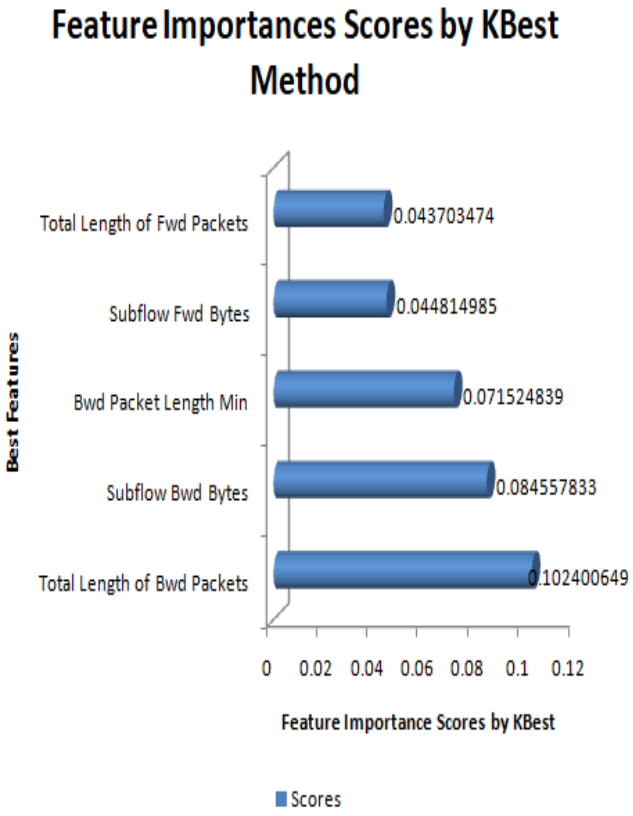

Fig. 7.Feature Importances of Best Features by KBest method

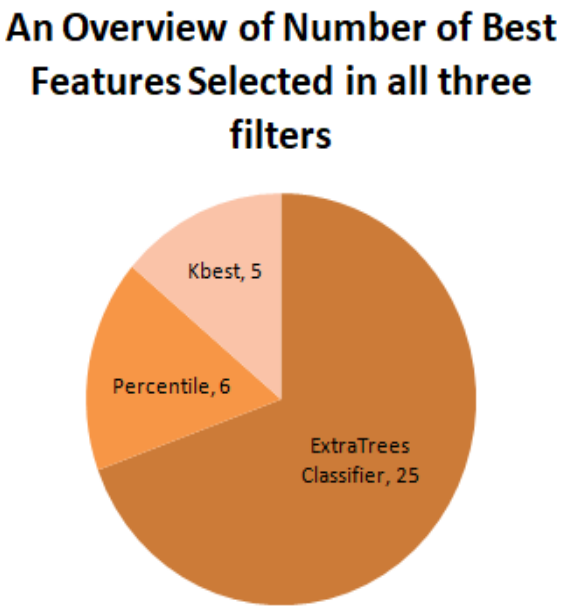

Fig. 8.The overall overview of feature selection by all three filters

Classification

Ensemble learning is mainly used to get an optimized solution. The combinations of several machine learning models or classifiers are formed by ensemble learning classifiers. The processed dataset FS is carried out for classification. There are 4 different ensemble learning classifiers AdaBoost, Gradient Boosting, Bagging and Random Forest are used for classification in the proposed work. Table-I shows the accuracy table for classification with the hybrid model and all three feature selection methods and with no feature selection methods. Among all the results, the Hybrid model with a random forest classifier gives the better result of 91.50 % accuracy.

V. RESULTS AND DISCUSSIONS

The results of the classification are taken under the conditions of 30 % testing data and 70% Training data as it is split from the whole dataset. The experimental results of the hybrid approach along with feature selection methods are given in Table-I and the respective figure of Table-I is represented in Fig. 9 which shows the performance metric as accuracy. Table-I shows that the Hybrid approach gives better results in all aspects of feature selection methods and ensemble classifiers. The Random Forest classifier gives better accuracy with 91.50%.

Table-I illustrates the results of all ensemble methods with a respective hybrid approach , feature selection methods and without using the feature selection method as tabulated. The accuracy of the classification model without using any feature selection methods are also tabulated in Table-I. All ensemble methods are given good results in hybrid approach. When compared hybrid approach with each feature selection methods ExtraTreesClassifier, Percentile and KBest , hybrid approach produce more better result. The last column of the table is listed that the accuracy of the classification of ensemble models to the original dataset without using any feature selection methods. The next among the Random Forest classifier the Bagging classifier gives the next better results with 91.04% accuracy.

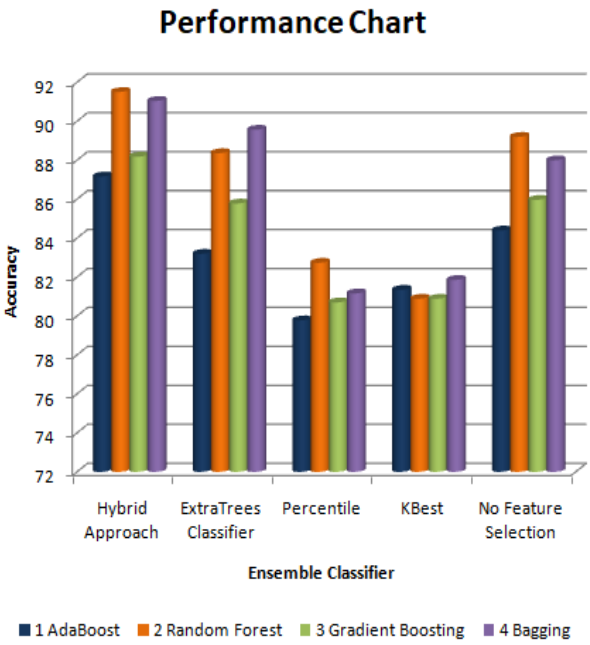

Fig. 9.Accuracy of classification model for ensemble classifiers

Table- I: Accuracy of Ensemble Classifiers with hybrid and other feature selection methods

S.No.	Ensemble Methods	Hybrid Approach	ExtraTrees Classifier	Percentile	KBest	No Feature Selection
1	AdaBoost	**87.1652**	83.1948	79.7783	81.3481	84.3951
2	Random Forest	**91.5050**	88.3656	82.7331	80.8864	89.1966
3	Gradient Boosting	**88.1809**	85.7802	80.7017	80.8864	85.9649
4	Bagging	**91.0434**	89.5660	81.1634	81.8559	87.9963

VI. CONCLUSION

The Hybrid approach of malware detection classification framework can effectively prove that it gives the best results of the classification of new malware samples and very much effective for high dimensionality of data. The ensemble classifier Random Forest gives the better result of 91.50% accuracy. The remaining classifiers are also given a better result when compared to the results of other feature selection methods and to the results of classification of original data i.e., without using feature selection methods. This hybrid framework works extremely when the high dimensionality of data. The filters method used in the hybrid framework can purify all the features and produce the most relevant best features basing on the score values or feature importance values. These filter methods can be applied more for the high dimensionality of data. The usage of different feature selection methods in filters instead of using single feature selection gives better results.

REFERENCES

1. Comparitech (2019). **Malware statistics and facts for 2019.** Available: https://www.comparitech.com/antivirus/malware-statistics-facts/
2. Liu, L., Wang, B. S., Yu, B., & Zhong, Q. X. (2017). Automatic malware classification and new malware detection using machine learning. *Frontiers of Information Technology & Electronic Engineering*, *18*(9), 1336-1347.
3. Gandotra, E., Bansal, D., & Sofat, S. (2014). Malware analysis and classification: A survey. *Journal of Information Security*, *5*(02), 56.
4. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28.
5. Khammas, B. M., Monemi, A., Bassi, J. S., Ismail, I., Nor, S. M., & Marsono, M. N. (2015). Feature selection and machine learning classification for malware detection. *Jurnal Teknologi*, *77*(1).
6. Yerima, S. Y., Sezer, S., & Muttik, I. (2015). High accuracy android malware detection using ensemble learning. *IET Information Security*, *9*(6), 313-320.
7. Mas' ud, M. Z., Sahib, S., Abdollah, M. F., Selamat, S. R., & Yusof, R. (2014, May). Analysis of features selection and machine learning classifier in android malware detection. In *2014 International Conference on Information Science & Applications (ICISA)* (pp. 1-5). IEEE.
8. Narudin, F. A., Feizollah, A., Anuar, N. B., & Gani, A. (2016). Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, *20*(1), 343-357.
9. Lu, Y. B., Din, S. C., Zheng, C. F., & Gao, B. J. (2010). Using multi-feature and classifier ensembles to improve malware detection. *Journal of CCIT*, *39*(2), 57-72.
10. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28.
11. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 94.
12. Mays, M., Drabinsky, N., & Brandle, S. (2017). Feature Selection for Malware Classification. In *MAICS* (pp. 165-170).
13. Scikit-learn (2019, December 20) [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html
14. Geeks for Geeks. (2019, December 20). A Computer Science Portal for Geeks [Online]. Available: https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/
15. Scikit-learn. (2019, December 20). [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectPercentile.html
16. Scikit-learn. (2019, December 20). [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html
17. Idrees, F., Rajarajan, M., Conti, M., Chen, T. M., & Rahulamathavan, Y. (2017). PIndroid: A novel Android malware detection system using ensemble learning methods. *Computers & Security*, *68*, 36-46.
18. Feng, P., Ma, J., Sun, C., Xu, X., & Ma, Y. (2018). A Novel Dynamic Android Malware Detection System With Ensemble Learning. *IEEE Access*, *6*, 30996-31011.
19. Alam, M. S., & Vuong, S. T. (2013, August). Random forest classification for detecting android malware. In *2013 IEEE international conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing* (pp. 663-669). IEEE.
20. Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

AUTHORS PROFILE

P. HarshaLatha, received her Bachelor's Degree in Computer Science from Sri Padmavathi Degree College, Tirupati, Andhra Pradesh, India in 2008. She received Post Gradute degree in Master of Computer Applications from Sreenivasa Institute of Technology and Management Studies, Chittoor, Andhra Pradesh, India in 2011. She has working experience as a Software Engineer in MNC Company from 2011 to 2013. She is currently pursuing M.Tech by Research as a Full time Research Scholar in the School of Computer Science and Engineering at Vellore Institure of Technology, Vellore, Tamilnadu, India. Her areas of interest are Machine Learning, Malware Analysis, Big Data, Cyber Security, Artificial Intelligence, Information Security.

Dr. Mohanasundaram Ranganathan, received his B.E. degree in Electrical & Electronics Engineering from Velalar College of Engineering and Technology, Erode in 2006, M.E. Embedded System Technologies from Velalar College of Engineering and Technology, Erode, Tamilnadu, India in 2008 and Ph.D. from Anna University, Chennai, India in 2015. He is currently as Associate Professor in the School of Computing Science and Engineering at Vellore Institute of Technology (Deemed University), Vellore, Tamilnadu, India. He has published more than 35 research papers in International and National Journals/Conferences. He have contributed few book chapters in the area of advanced embedded systems and swarm intelligence. His areas of interest are Swarm Intelligence, Wireless Sensor Networks, Computer Networks, VLSI, Embedded Systems, Mobile Communications, Life science.