SEYBOLD Report
ISSN: 1533 - 9211
https://seyboldreport.net/

# FEATURE SELECTION IN MACHINE LEARNING

**Meenakshi[1], Ramachandra AC[2]**
[1]Research Scholar, Computer Science & Engineering Department,
Nitte Meenakshi Institute of Technology, Bangalore, India. meenakshi.rao.kateel@gmail.com
[2]Professor & Head, Electronics & Communication Engineering Department,
Nitte Meenakshi Institute of Technology, Bangalore, India, ramachandra.ac@nmit.ac.in

**Abstract**
*In this paper we discuss importance of selecting features and various approaches for feature selection during Machine Learning modeling. Data cleaning and Feature selection are two important steps carried out in the beginning of every machine learning model designing. While final features selected for modeling have a significant effect on performance, on the other side insignificant features give a negative effect of performance. For Deep Learning models, explicit feature selection is not advisable due to presence of its inbuilt internal feature selection while modelling itself. Depending on domain and dataset, still feature selection can be applicable for reducing time and space complexity. It also helps to explore weak, noisy, irrelevant features present in the dataset.*
**Keywords:** Deep Learning, Feature Selection, Machine Learning, Supervised, Unsupervised

## I. INTRODUCTION

Objective of feature selection is to bring down the input variables to "those that are believed to be most useful to a prediction model". Less significant features are considered as noise which lead to less accurate models. There are many advantages of selecting best suitable features contributing to prediction. Although it is not mandatory for deep learning models also, we can apply feature selection. It serves the purpose of exploration and domain specific feature selection.[1]

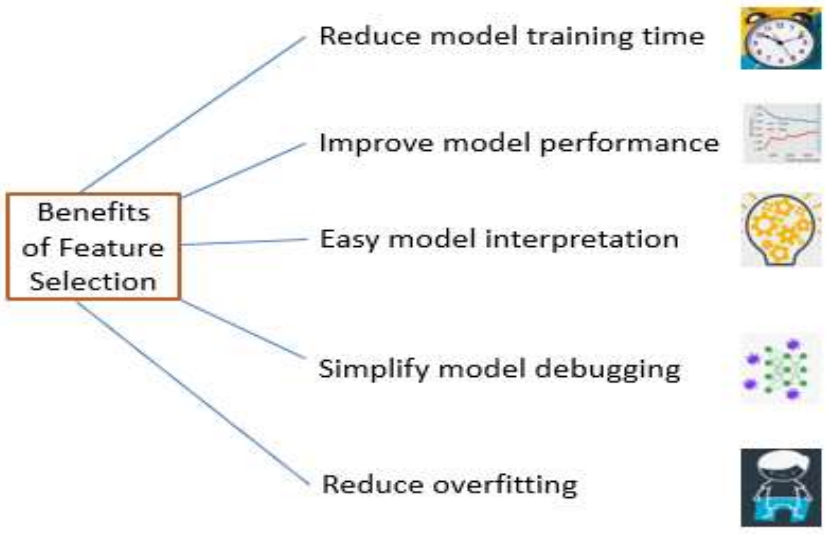

Figure 1: Advantages of Feature Selection

Supervised and unsupervised are two approaches of eliminating features. In supervised, target variable is considered while selecting features and in unsupervised target variable is ignored. [2-4]

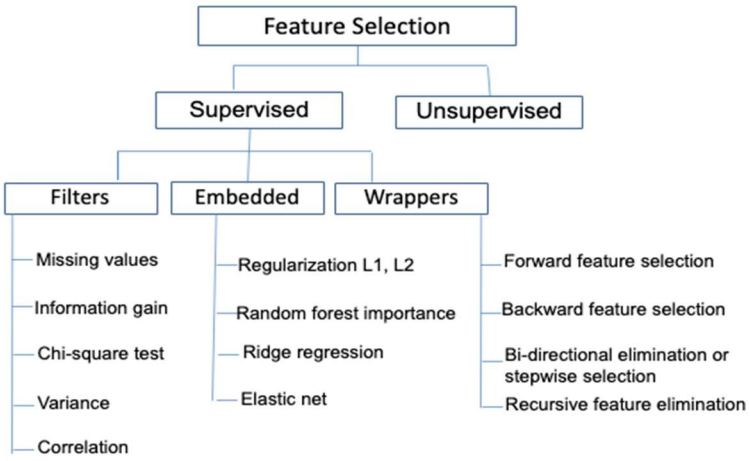


Figure 2: Feature Selection Methods

| | Filter method | Wrapper method | Embedded method |
|---|---|---|---|
| 1 | Uses proxy measures such as correlation | Uses predictive model | Feature selection method is embedded in the model building phase |
| 2 | Computationally faster | Slower | Medium |
| 3 | Avoids overfitting | Vulnerable to overfitting | Less vulnerable to overfitting |
| 4 | Sometimes may fail to select best features | Better performance | Good performance |
| 5 | Less expensive | Computationally very expensive | Moderate |

Table 1: Comparison of Feature Selection methods [5-10]

## FILTER METHOD

In this approach, based on some statistics calculations, ranking methods are applied. Features are filtered by ordering. It is a part of pre-processing stage and independent of machine learning algorithm. Various statistical tests are performed on features and scores obtained on correlation with output feature is used as criteria for feature selection. Some of the important Filter methods are depicted below.

**Missing values** - Although imputation is possible for missing values, it is good approach to drop columns which have missing values greater than some predefined threshold. Features with large number of missing values may not be very significant. [11] et al. discussed "risk/benefit" trade-off happens due to assigning missing values. Some of the similar work is done in [12] and [13].

**Information Gain -** It is used for decision tree construction and for feature selection. It is referred as *mutual information* when it is used for feature selection. Every independent variable's gain is calculated in connection with target variable. It is also referred as calculating "statistical dependency" among independent and dependent variables.
It is stated mathematically as below:

$$I(X; Y) = H(X) - H(X \mid Y),$$

where X and Y are random features. "H(X) is the entropy for X and H (X | Y) is the conditional entropy for X given Y". The result has the units of bits. "It measures the average reduction in uncertainty about x that results from learning the value of y; or vice versa, the average amount of information that x conveys about y". Information gain will be greater than or equal to 0. Here 0, it indicates X and Y are independent. Larger value indicates more dependency among X and Y. [14-16]

**Chi-Square Test-** It is the "measure of association between two categorical variables". It helps to address, how two categorical variables dependent on each other. Chi-square test assumes that observed and expected frequencies for a categorical value remains same. It is called Null Hypothesis. Formula for Chi-square test is given below:

$$\chi^2 = \sum (O_i - E_i)^2/E_i$$

where $O_i$ is "observed frequency" and $E_i$ is "expected frequency" for the categorical feature. "The variables are considered independent if the observed and expected frequencies are similar, that the levels of the variables do not interact, are not dependent". [17-19]

**Variance Threshold**- In sample set if a variable shows less variance, it is assumed to be less significant towards determining dependent feature. Using some threshold, such variables are removed before modeling. Problem here is, relationship to dependent feature or with other independent features is not considered. [20-21]

**Correlation Coefficient**- Correlation is "measure of strength" between two continuous variables. "It is a statistical term which in common usage refers to how close two variables are to having a linear relationship with each other". Score of independent variables is used to filter irrelevant features. Score is obtained from various statistical tests where correlation of independent features with dependent feature is obtained. High correlation among independent features indicates there exists linear dependency among them. Hence both features have almost

same impact on dependent feature also. [22-23]

**WRAPPERS**

It means wrapping features. Specific machine learning algorithm is chosen to fit the dataset for feature selection. Various combinations of dependent variables are used to produce multiple models and subset of features which yielded best performing model is selected as final feature set. It is a greedy approach because all possible subset of features is used, evaluated against evaluation criteria such as accuracy, precision, recall etc. in case of classification and p-values, R-squared etc. in regression. Important Wrapper techniques are given below. [24-27]

**Forward Selection-** It starts with empty feature set and in every iteration new feature which improves performance better is added to this set. This process continues until adding of new feature will not improve the performance further.

**Backward Elimination-** It starts with all given features and then in each iteration we eliminate least significant feature which improves performance further until we reach the stage where removal will not contribute to performance.

**Bi-directional Elimination or Stepwise Selection-** It is combination of forward and backward feature selection. It works like forward selection and while adding new feature it also removes less significant feature from already selected set.

**Recursive Feature Elimination-** Most relevancy of feature towards predicting target is main key point here. It gives flexibility of selecting how many top features are to be selected and which ML algorithm to be used internally for feature selection. This core algorithm is wrapped by RFE. It is greedy in nature.

**EMBEDDED APPROACH**

It makes use of properties of both filter and wrapper methods. Algorithms have their own built-in variable selection methods. [28-30]

**Random Forest Importance-** Based on tree decision, it is most effective to compute feature importance towards decision. The word "random" refers to random picking of observations and features. It simply implies that out of hundreds of trees, each one represents random set of rows and random set of columns from dataset. Due to random selection, it leads to less correlation among trees. Each random tree, gives two separate buckets of observations which are similar with in the bucket and almost completely different than other bucket. From all random decision trees, constructive result is taken for final feature selection.

## II. IMPLEMENTATION

Data set used for experiment is CIC-DDoS-2019. It has 82 features. We have implemented the following feature selection methods.

**Information Gain**

It is an example of *Filter* method of feature selection. We measure information gain among dependent and independent variable and more the information gain, more the dependency between dependent and independent variable. Information gain calculation is based on "entropy estimation" from KNN distances.

| **Algorithm** | |
|---|---|
| Step1 | Import mutual_info_classif from sklearn.feature_selection library |
| Step2 | Prepare datasets |
| | X <-- independent features |
| | y <-- dependent feature 'Label' |
| Step3 | Fit the model |
| | mutual_info <-- mutual_info_classif(X, y) |
| Step4 | Sort it in ascending order |
| | Select the top N features |
| Step5 | If *matrix* value is above threshold mark array entry as *False for that column* |
| Step6 | Repeat step 5 for all entries in *correlation matrix* |

Table 2: Information Gain Algorithm

| # | **Independent Feature** | **Score** |
|---|---|---|
| 1 | Destination Port | 0.251857 |
| 2 | Flow Packets/s | 0.205207 |
| 3 | Bwd Packets/s | 0.202705 |
| 4 | Flow IAT Mean | 0.198202 |
| 5 | Max Packet Length | 0.197924 |
| 6 | Flow IAT Max | 0.197907 |
| 7 | Inbound | 0.197836 |
| 8 | Fwd Packet Length Mean | 0.194803 |
| 9 | Fwd Packets/s | 0.194394 |
| 10 | Avg Fwd Segment Size | 0.193659 |
| 11 | Average Packet Size | 0.193353 |
| 12 | Packet Length Mean | 0.191692 |
| 13 | Fwd Packet Length Max | 0.187157 |
| 14 | Source Port | 0.186707 |
| 15 | Subflow Fwd Bytes | 0.184240 |
| 16 | Total Length of Fwd Packets | 0.183234 |
| 17 | Flow Duration | 0.177495 |
| 18 | Flow Bytes/s | 0.172031 |

| | | |
|---|---|---|
| 19 | Flow IAT Std | 0.170881 |
| 20 | Fwd Packet Length Min | 0.168864 |

Table 2: Top 20 Features with their Importance obtained from Information Gain

**Feature Selection based on Correlation among other features:**

This is another implementation example for *Filter* method of feature selection. Sometimes independent features exhibit multicollinearity where there exists strong association with each other. Heatmap gives correlation coefficient values of all feature pairs. Basic idea is to "remove highly correlated features". If input features are highly correlated with output variable, then need not to remove them. If input features are highly correlated among themselves then need to remove them and train the model with remaining features. When one independent feature is highly correlated with another independent feature means one can be predicted from other. In that scenario, model needs any one of them. Correlation matrix helps to select or drop features.

| **Algorithm** | |
|---|---|
| Step1 | Set *threshold* value to filter features |
| Step2 | Create *dataframe* by excluding dependent feature |
| Step3 | Derive the *correlation matrix* by *corr*() function |
| Step4 | Create an array of size as features fill with *True* |
| Step5 | If *matrix* value is above threshold mark array entry as *False for that column* |
| Step6 | Repeat step 5 for all entries in *correlation matrix* |

Table 3: Correlation based feature selection – An Algorithm

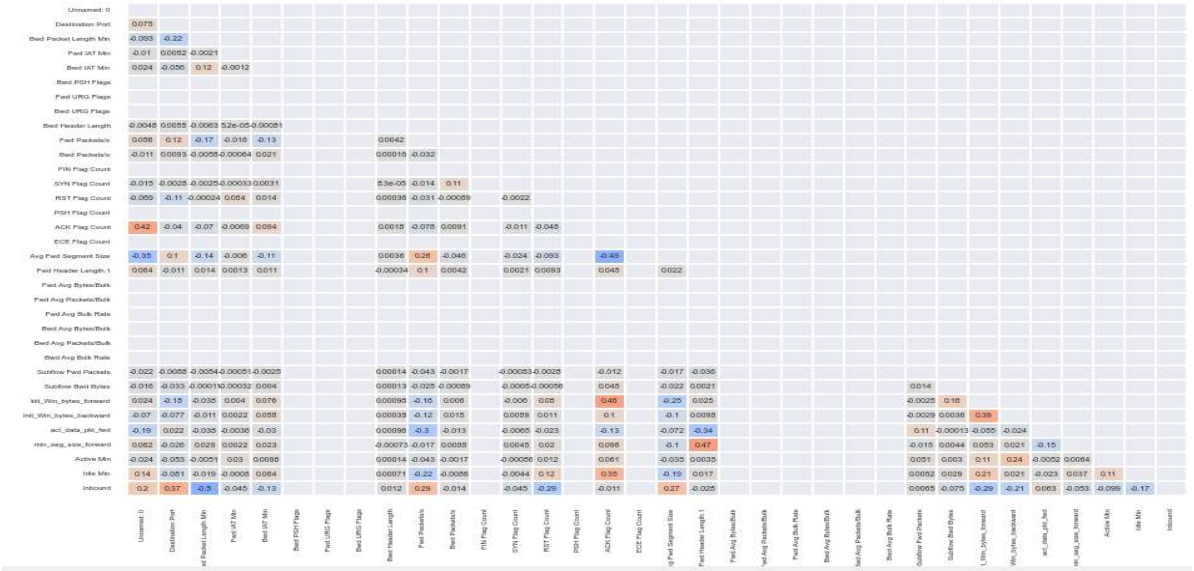

Figure 3: A correlation Heatmap

Heatmap in the Figure 3 shows the correlation coefficients among features. Since we excluded

dependent feature, heat map contains relationship among independent features only. If this correlation coefficient or dependency is above the given threshold then we select one of those features and drop the another.

Top 20 Features selected based on Correlation Coefficient are: "Source Port, Destination Port, Bwd Packet Length Min, Fwd IAT Min, Bwd IAT Max, Bwd IAT Min, Bwd PSH Flags, Fwd URG Flags, Bwd URG Flags, Bwd Header Length, Fwd Packets/s, Bwd Packets/s, FIN Flag Count, SYN Flag Count, RST Flag Count, PSH Flag Count, ECE Flag Count, Down/Up Ratio, Avg Fwd Segment Size, Fwd Header Length.1, Fwd Avg Bytes/Bulk, Fwd Avg Packets/Bulk, Fwd Avg Bulk Rate, Bwd Avg Bytes/Bulk, Bwd Avg Packets/Bulk, Bwd Avg Bulk Rate, Subflow Fwd Packets, Subflow Bwd Bytes, Init_Win_bytes_forward, Init_Win_bytes_backward, act_data_pkt_fwd, min_seg_size_forward, Active Min, Idle Min and Inbound".

**Feature Selection based on RFCV (RFE Cross Validation):**

It is an example for *Wrapper* based feature selection.

| **Algorithm** | |
|---|---|
| **Step1** | Set *estimator* algorithm for feature selection |
| **Step2** | Fit model to dataset |
| **Step3** | Eliminate feature with smallest coefficient or lowest ranking feature. |
| **Step4** | Repeat Step 2 and 3 until elimination will not improve the performance of the model. |

Table 3: RFE working Methodology

In RFE, number of features to be selected is mentioned. In RFECV no need to mention how many features are required explicitly, it only computes and gives final list of features. The algorithm chosen for *estimator* should be able to important scores such as "feature importance" in case of decision trees or "coefficients" in case of linear regression. Lesser the coefficient lesser the importance. "*DecisionTreeClassifier"* is used here as *estimator.*

Top ranked 20 features selected are: "Source Port, Destination Port, Total Fwd Packets, Total Backward Packets, Total Length of Fwd Packets, Total Length of Bwd Packets, Fwd Packet Length Min, Flow Bytes/s, Flow Packets/s, Fwd Packets/s, Min Packet Length, Average Packet Size, Avg Fwd Segment Size, Avg Bwd Segment Size, Fwd Header Length.1, Init_Win_bytes_forward, act_data_pkt_fwd, Active Min, Idle Min, Inbound.

**Feature Importance based Feature Selection**

This is one of the *Embedded* categories of feature selection also known as E*xtremely Random Tree Classifier* or Extra Tree Classifier (ETC). This technique gives score of each feature and more the score more relevant the feature for prediction. Table4 gives the list of top features selected based on feature importance algorithm. Figure4 is the visualization of the features selected.

| # | Independent Feature | Score |
|---|---|---|
| 1 | Inbound | 0.304938 |
| 2 | URG Flag Count | 0.080809 |
| 3 | CWE Flag Count | 0.045347 |
| 4 | ACK Flag Count | 0.042008 |
| 5 | Destination Port | 0.039658 |
| 6 | min_seg_size_forward | 0.039025 |
| 7 | Source Port | 0.037485 |
| 8 | Fwd Packet Length Min | 0.027226 |
| 9 | Min Packet Length | 0.025957 |
| 10 | Bwd Packet Length Min | 0.024127 |
| 11 | Fwd Packets/s | 0.022563 |
| 12 | Init_Win_bytes_forward | 0.022459 |
| 13 | Down/Up Ratio | 0.021411 |
| 14 | Avg Fwd Segment Size | 0.019847 |
| 15 | Protocol | 0.018201 |
| 16 | Fwd Packet Length Mean | 0.017925 |
| 17 | RST Flag Count | 0.015059 |
| 18 | Fwd PSH Flags | 0.014029 |
| 19 | Average Packet Size | 0.012642 |
| 20 | Init_Win_bytes_backward | 0.012409 |

Table 4: Top 20 Features and their Importance obtained from ETC

ETC makes use of uncorrelated decision trees at the backend. It uses all records in the sample for constructing every tree. In each node it selects random parameters for split. Whereas Random Forest selects best available parameters at each node for split based on Gini or Entropy. It is greedy algorithm.

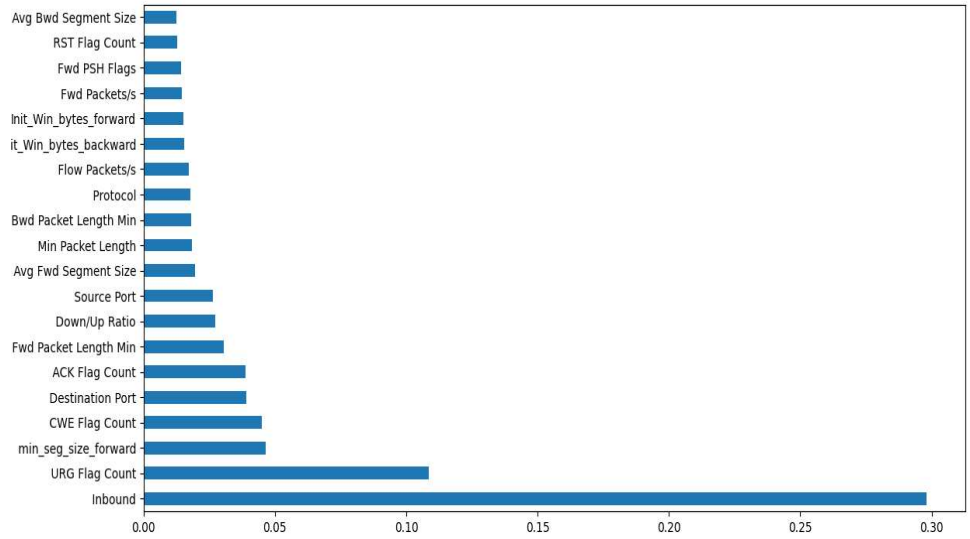

Figure 4: Feature Importance based Feature Selection

## III. Conclusion

Feature selection is one of the important steps in Prediction Model building. We presented comprehensive methodologies of feature selections categories and types. CIC-DDoS 2019 dataset is taken as a benchmark for all our experiments. Survey and experimental results presented can become base reference for exploring domain specific feature selections.

## IV. Reference

[1] Guilherme Perin, Lichao Wu, Stjepan Picek, "Exploring Feature Selection Scenarios for Deep Learning-based Side-channel Analysis", Cryptology ePrint Archive, 2021. https://eprint.iacr.org/2021/1414

[2] KenjiKira, Larry A.Rendell, "A Practical Approach to Feature Selection", Machine Learning Proceedings, 1992. https://doi.org/10.1016/B978-1-55860-247-2.50037-1

[3] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE, vol. 17, no. 4, pp. 491–502, 2005.

[4] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics", Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.

[5] Samina Khalid, Tehmina Khalil and Shamila Nasreen, "A survey of feature selection and feature extraction techniques in machine learning Publisher: IEEE Cite This PDF", Science and Information IEEE Conference, 2014. DOI: 10.1109/SAI.2014.6918213

[6] Bing Xue, Mengjie Zhang, Will N. Browne and Xin Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection", IEEE Transactions on Evolutionary Computation, vol. 20, 2016. DOI:10.1109/TEVC.2015.2504420

[7] Cai, Jie, Luo, Jiawei, Wang, Shulin, Yang, Sheng, "Feature selection in machine learning: A new perspective", Neurocomputing, vol. 300, 2018. https://doi.org/10.1016/j.neucom.2017.11.077

[8] M. Sharma and P. Kaur, "A comprehensive analysis of nature-inspired meta-heuristic

techniques for feature selection problem", Arch. Comput. Methods Eng., pp. 1–25, 2020. DOI: 10.1007/s11831-020-09412-6.
[9] P. Agrawal, H. F. Abutarboush, T. Ganesh and A. W. Mohamed, "Metaheuristic Algorithms on Feature Selection", vol. 9, IEEE, 2021. DOI: 10.1109/ACCESS.2021.3056407
[10] M. Dash, H. Liu, "Feature selection for classification", Intelligent Data Analysis, vol. 1, Issues 1–4,1997, ISSN 1088-467X. https://doi.org/10.1016/S1088-467X(97)00008-5.
[11] Borja Seijo-Pardo, Amparo Alonso-Betanzos, Kristin P. Bennett, Verónica Bolón-Canedo, Julie Josse, Mehreen Saeed, Isabelle Guyon, "Biases in feature selection with missing data", Neurocomputing, vol. 342, 2019, pp. 97-112, ISSN 0925-2312. https://doi.org/10.1016/j.neucom.2018.10.085.
[12] Misra, Puneet and Yadav, Arun Singh, "Impact of Preprocessing Methods on Healthcare Predictions", Advanced Computing and Software Engineering 2019. http://dx.doi.org/10.2139/ssrn.3349586
[13] Robert Tibshirani, "The Lasso Method for Variable Selection In The Cox Model", vol. 16, Issue 4, 1997. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3
[14] L. F. Kozachenko, N. N. Leonenko, "Sample Estimate of the Entropy of a Random Vector", Probl. Peredachi Inf., 23:2, 1987, 9-16
[15] A. Kraskov, H. Stogbauer and P. Grassberger, "Estimating mutual information". Phys. Rev. E 69, 2004.
[16] B. C. Ross "Mutual Information between Discrete and Continuous Data Sets". PLoS ONE 9(2), 2014.
[17] Lynne Connelly, "Chi-Square Test", Medsurg Nursing, 2019. https://www.proquest.com/openview/04d2ff080887f9111b68eb7490a9630a/1?pq-origsite=gscholar&cbl=30764
[18] Said Bahassine, Abdellah Madani, Mohammed Al-Sarem, Mohamed Kissi, "Feature selection using an improved Chi-square for Arabic text classification", Journal of King Saud University - Computer and Information Sciences, vol. 32, Issue 2, 2020. https://doi.org/10.1016/j.jksuci.2018.05.010.
[19] Pattepu Naresh, Teja Tallam, Naveen Kumar, Chikkakrishna, "Analysis of Urban Traffic Bottleneck in Hyderabad city using Machine Learning Techniques", International Conference on IoT in Social, Mobile, Analytics and Cloud, 2020. DOI: 10.1109/I-SMAC49090.2020.9243308
[20] Sánchez-Maroño, N., Caamaño-Fernández, M., Castillo, E., Alonso-Betanzos, A. "Functional Networks and Analysis of Variance for Feature Selection", Intelligent Data Engineering and Automated Learning, 2006, Lecture Notes in Computer Science, vol. 4224. Springer, https://doi.org/10.1007/11875581_123
[21] M. Al Fatih Abil Fida, T. Ahmad and M. Ntahobari, "Variance Threshold as Early Screening to Boruta Feature Selection for Intrusion Detection System," Information & Communication Technology and System, 2021, pp. 46-50, DOI: 10.1109/ICTS52701.2021.9608852

[22] Hall, M. A. & Smith, L. A, "Practical feature subset selection for machine learning", Australasian Computer Science Springer Conference Perth, 1998, pp. 181-191. https://hdl.handle.net/10289/1512
[23] A. Kaur, K. Guleria and N. Kumar Trivedi, "Feature Selection in Machine Learning: Methods and Comparison", Advance Computing and Innovative Technologies in Engineering 2021, pp. 789-795, DOI: 10.1109/ICACITE51222.2021.9404623.
[24] S. Cohen, G. Dror and E. Ruppin, "Feature Selection via Coalitional Game Theory", Neural Computation, vol. 19, pp. 1939-1961, 2007, DOI: 10.1162/neco.2007.19.7.1939.
[25] Girish Chandrashekar, Ferat Sahin, "A survey on feature selection methods", Computers & Electrical Engineering, vol.40, Issue 1, 2014. https://doi.org/10.1016/j.compeleceng.2013.11.024.
[26] Rani, Pooja,et al, "A Hybrid Approach for Feature Selection Based on Genetic Algorithm and Recursive Feature Elimination", vol.12, pp.17-38. 2021. http://doi.org/10.4018/IJISMD.2021040102.
[27] Neha V Sharma, Narendra Singh Yadav, "An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers, Microprocessors and Microsystem", vol. 85,2021. https://doi.org/10.1016/j.micpro.2021.104293.
[28] Osanaiye, O., Cai, H., Choo, KK.R. et al. "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing", Wireless Com Network, 2016. https://doi.org/10.1186/s13638-016-0623-3
[29] S. Das, D. Venugopal, S. Shiva and F. T. Sheldon, "Empirical Evaluation of the Ensemble Framework for Feature Selection in DDoS Attack", Cyber Security and Cloud Computing, 2020, pp. 56-61, DOI: 10.1109/CSCloud-EdgeCom49738.2020.00019.
[30] Nivedhitha Mahendran, Durai Raj Vincent, "A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease", Computers in Biology and Medicine, vol. 141, pp. 105056, 2022. DOI: https://doi.org/10.1016/j.compbiomed.2021.105056