Third International Conference on "Materials, Computing and Communication Technologies"
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

# Exploring Various Intrusion Detection Methods Using Machine Learning Techniques

**Archana R. Ugale [1], Dr. Amol Potgantwar[2]**

[1]Research Scholar, School of Engineering & Technology,D.Y.Patil University, Ambi Pune, Maharashtra, India
[2]Associate Professor, Department of Computer Engineering, Sandip Institute of Technology & research Center, Nashik, Maharashtra, India

**Article Info**
Volume 8, Issue 7
Page Number : 236-242

**Publication Issue :**
May-June-2022

**Article History**
Accepted: 01 June 2022
Published: 20 June 2022

**ABSTRACT**

In the modern world of security, it is very necessary to put in place intrusion detection systems (IDS) and intrusion prevention systems (IPS) that are both reliable and effective. The primary function of an intrusion detection system (IDS) is to identify unusual behavior in network traffic by making use of efficient techniques. In intrusion detection based on anomaly detection, the application of machine learning algorithms plays an essential part. The purpose of this study is to provide an overview of the machine learning approaches that are utilized in intrusion detection. For intrusion detection, classification techniques such as logistic regression, naive Bayes, KNN, Decision Tree, Random Forest, and Support Vector Machines (SVM) work well. In this research, we investigate the behavior and features of classification approaches that are based on machine learning and are utilized for applications related to intrusion detection.

**Keywords** :— Intrusion Detection System, Intrusion Prevention Systems, Machine Learning

## I. INTRODUCTION

Intrusion is the process of breaking into a secure network in an unauthorised manner with the intention of committing a malicious act such as stealing information or posing a risk to the server or information system [8]. Anomaly detection was the primary method used in intrusion prevention and detection. An intrusion may have occurred due to a strange occurrence in the network flow. In addition, it is checked, and permission is granted, because there is always the possibility of a false positive. The difference between an intrusion detection system and an intrusion prevention system is that the former simply reports when an intrusion has occurred, while the latter remembers the pattern of the intrusion and helps to prevent such incidents in the future. IPS is accountable for preventing the attack on its own and recording the incursion for the purpose of providing additional protection in the future [9]. Techniques based on machine learning play a significant part in the process of intrusion detection. The classification strategies that are a part of machine learning come in handy when trying to identify an intrusion [10], [12].

The majority of the time, an anomaly that is discovered in the traffic on the network is interpreted as an intrusion. Techniques for detecting outliers or anomalies play their own unique part in the intrusion detection process. The techniques of Logistic regression, Naive Bayes, RNN, Decision tree, Random Forest, and Support Vector Machines (SVM) are discussed in detail throughout this work as they pertain to intrusion detection. Logistic regression is the statistical model that is utilised in machine learning for the purpose of performing classification tasks. The sigmoid function provides the foundation for the work that is done in logistic regression [13].

The Bayes theorem is utilised to classify the data using the naive Bayes classification model, which is a probability-based classification model. This model is utilised to do classification on the dataset. The well-known classification model known as K-Nearest Neighbour performs classification based on three different kinds of distance known as Euclidean, Manhattan, and Minkowski[11]. In the field of machine learning, a well-known type of categorization model is called a decision tree. Each internal node of the decision tree is responsible for processing one attribute before contributing to the final conclusion. Classes are denoted by the leaf nodes. The random forest algorithm is an extension of the decision tree algorithm. The random forest algorithm is the combining of a large number of decision trees in order to increase the performance of categorization. A hyperplane in a two-dimensional space divides the data into two groups, which is the classification that the support vector machine algorithms arrive at when applying themselves to the data. The accuracy of the categorization can be determined by the size of the margin that surrounds the hyperplane. In this research, we analyse the work that was done before with regard to detecting intrusions in a variety of contexts. The fuzzy C-means and logistic regression-based intrusion detection system was presented for the cloud environment by Pkanomozhi et all., 2021. The method based on logistic regression is utilised for feature selection, and the algorithm based on fuzzy C-means clustering is used to cluster normal behaviour and malignant activity [1].[7].

The research of Pan, J. S., et al.,2021 Combining the k-nearest neighbour algorithm (kNN) and the sine cosine algorithm (SCA) is an idea that has been proposed for use in the detection of intrusions in wireless sensor networks. The temporal complexity issue can be resolved with the intrusion detection approach that makes use of the k-nearest neighbour and sine cosine algorithm [2]. To cite this article: Panigrahi, R., et al.,2021 It was suggested that the decision tree-based C4.5 algorithm and the consolidated tree construction (CTC) method should be combined in order to achieve a higher level of accuracy in intrusion detection. This approach of intrusion detection does not focus on a particular environment; rather, it focuses on an imbalanced dataset. The suggested method addresses the problems associated with an imbalanced dataset by employing preprocessing techniques [Refppr3]. Chen, Z., et al.,2021 [Chein, Z., et al.,2021] The random forest method was proposed as a basis for the intrusion detection. An method based on a random forest is utilised here in order to increase the performance of the technique for detecting intrusions. Metrics known as precision, recall, and the F1 score are used to evaluate the effectiveness of the random forest method [4].

Gu, J., et al.,2021 [Gu et al.] A method for the identification of intrusions using naive Bayes and support vector machines was proposed. This method of intrusion detection uses naive Bayes, which is used to the actual data after the actual data have been preprocessed and transformed into quality data. In order to identify the suspicious activity, an SVM analysis is performed on the converted data. The elimination of erroneous "positive" detections is the primary goal of both the support vector machine and the naive Bayes technique [5]. Kurniawan, Y. I., and all of 2021, etc. For the purpose of intrusion detection, the modified naive Bayes method was proposed. When it comes to dealing with zero probability, the current version of the naive Bayes algorithm has some issues. In the technique that has been proposed, the issue at hand can be resolved by making alterations to the formula. The precision, recall, F-score, and accuracy measures have all been utilised in order to evaluate the accuracy of the model [6].

## II. RELATED WORK

### A. Logistic Regression-based Intrusion Detection using Fuzzy C-Means

It is an unsupervised clustering approach that was created from the K-means clustering algorithm. Fuzzy C-means is a type of clustering algorithm. When using the K-means clustering algorithm, each and every data object is given a cluster to belong to based on the distance that exists between the data object and the centroid of the cluster. Every data object maintains a fuzzy-based membership value in order to take part in the fuzzy Cmeans algorithm. This allows the data objects to participate in each cluster. The fuzzy C means clustering formula to calculate the centroid is distinct from the k-means clustering technique in its operation.

The most common application of the probability-based statistical model known as logistic regression is in the process of binary categorization. The logistic regression method also incorporates aspects of the linear regression method. Any value in the range from 0 to infinity will be produced as the output of the linear regression component. In order to complete the classification using linear regression, it is necessary to restrict the range of possible values. The linear regression output value is passed to the sigmoid function, and the range is confined between 0 and 1 in order to achieve the desired effect of limiting the output's possible values.

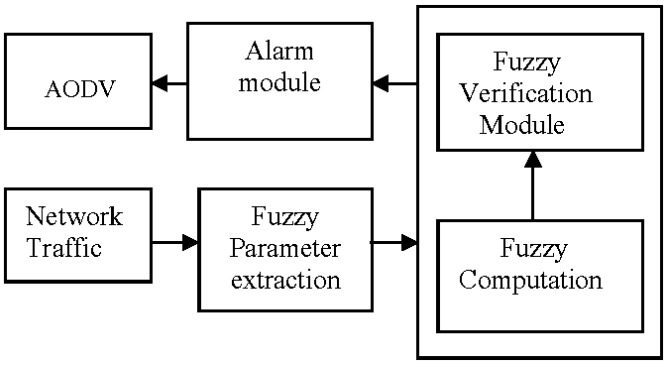


Fig 2.1: Fuzzy Logic based IDS

The number 0.5 is assumed to be the threshold limit for the tuples; those tuples whose values are lower than 0.5 are classified as belonging to class 1, while those with values greater than 0.5 are classified as belonging to class 2. The method of intrusion detection approach was discovered for the cloud environment in the method that was proposed by P.kanimozhi et all.,2021. [1]. In order to detect the intrusion in the cloud environment, it suggested a technology known as logistic regression-based oppositional truncate fuzzy Cmeans (LR-OTSFCM). The selection of features is the primary function of logistic regression, which has been carried out as a component of preprocessing. In this stage of preprocessing, feature selection is carried out not by dimensionality reduction but rather using logistic regression.

### B. KNN-based Intrusion Detection

For classification purposes, the k-nearest neighbour method is a popular choice among machine learning practitioners. Despite the fact that KNN can also be used to handle problems involving regression, it is most commonly employed for classification. In educational parlance, this is referred to as "supervised learning." When there is a new data object to classify, the data objects are represented in n-dimensional space based on the training data, and k-nearest data objects are evaluated surrounding the new data object depending on the value k.

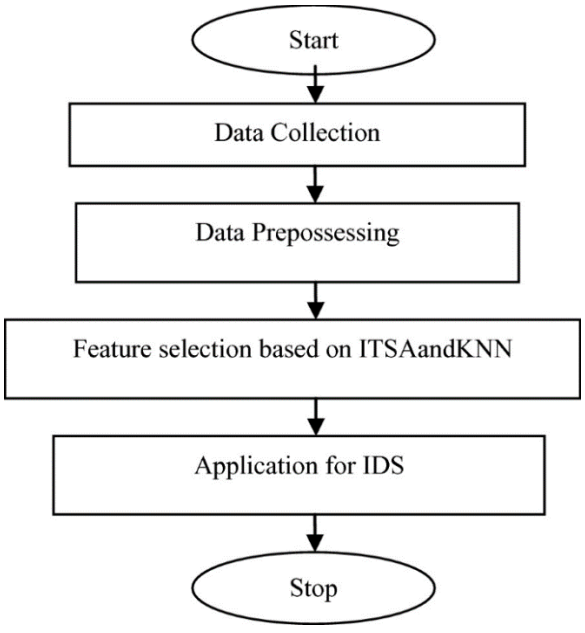


Fig 2.2: KNN based IDS

This process continues until the new data object has been classified. After determining the count of data

objects that belong to each class in the k-nearest neighbour set, it is determined which class will store the data item based on which class has the highest count. The classes of all of the data objects that make up the K-nearest neighbour set are already well known. The class with the most immediate neighbours for the newly created data is the one that gets assigned to the object representing the newly created data. The classification is being place in this manner using KNN. Because of its ease of use and widespread adoption, KNN has become a well-known and popular categorization method. Intrusion detection using KNN and sine cosine algorithm was proposed by Pan, J. S., et all.,2021 [2] in order to identify the outlier in a wireless sensor network. The accuracy of intrusion detection has been significantly improved because to the utilisation of KNN. The number of false-positive results produced by the approach saw a significant reduction as a result of the changes.

### C. Intrusion detection using a decision tree

Another well-known categorization strategy that may be applied using machine learning is the decision tree. The tree is built using the training data set as its foundation. The non-leaf nodes of the tree are given conditions based on the condition that the tree is split into branches, with each branch storing a subset of the training data set. After the construction of the decision tree, the next step is the classification of the new data object. After the construction of the decision tree has been successfully concluded, it is simple to classify newly acquired data objects. The decision tree method functions in a manner that is quite at ease with the high dimensional data set. The decision tree is used for intrusion detection with the primary goal of increasing the level of accuracy achieved. The measured entropy is used to choose which characteristic should be tried in the root note and which should be examined in the other non-leaf notes. The value of the entropy ranges between 0 and 1, and it is this range that is used to make the decision. A lower entropy is usually preferable because it results in a shorter and more manageable tree by reducing the height of the tree. All of the nodes that are not leaf nodes have conditions, while leaf nodes have classes. The attribute that has the lowest entropy occupies the root node, and other attributes maintain their position in the tree according to their entropy value.

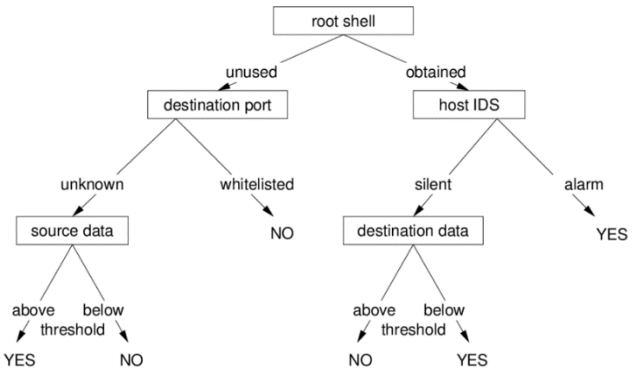


Fig 2.3: Decision Tree based IDS

A strategy for detecting intrusions using the decision tree was proposed by Panigrahi, R., et al., 2021 [3] for the multi-class imbalanced dataset. The unbalanced data set has a tendency to lean in one particular direction. In a balanced data set, the majority of the classes have the same number of tuples, while in an imbalanced data set, the majority of the tuples belong to just one class, while the other classes have a significantly lower total number of tuples. Intrusions are not typical activities; rather, they are considered abnormal activities. When we monitor the data for signs of intrusion, we find very few positive samples, while there are a greater number of suspicious negative samples; as a result, the data associated with intrusion detection is imbalanced. The approach of intrusion detection that has been developed uses an algorithm that is based on C 4.5. During the pre-processing stage, samples are collected according to class for analysis. As part of the pre-processing step, features are chosen according to the technique of choice for feature selection. Once the data has been prepared for intrusion detection by having a sufficient number of features and samples, it is ready for the process. At this point, decision tree-based detection is carried out on the data.

### D. Intrusion detection system using Random Forest algorithm

An algorithm that is based on decision trees has been extended in order to create the Random Forest algorithm. The decision tree is a classification

algorithm that is used to select one class out of n number of classes for the data that has been given. The random forest algorithm combines n number of decision trees[15], the outcome of every decision tree is taken into consideration, and the class selection is based on the decision of the maximum number of decision trees. Following in the footsteps of ensemble learning is the random forest algorithm. The strategy known as "ensemble learning" is one in which numerous different classifiers are combined into a single model in order to more effectively address the problem. The subsets of the provided data are each given their own decision tree, and the outcome of the random forest is a composite of the results of each decision tree. The ultimate class is assigned according to the decision reached by the majority of the decision trees.

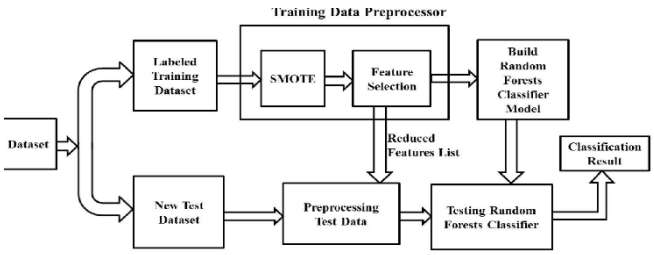


Fig 2.4: Random Forest based IDS

Using a random forest algorithm that was trained with the CICIDS 2017 dataset, Chen, Z., et al., (2021) [4] suggested a method for intrusion detection. The dataset was brought into equilibrium with the help of adaptive synthetic sampling. The use of random forests for intrusion detection yielded positive results across a broad spectrum of datasets. For the purpose of sampling, the stratified k-fold sampling approach is utilised, and after that, an intrusion detection algorithm based on a random forest classifier is applied to the sample data. The flow diagram of the rando forest classifier can be seen in Fig 2.1, and the work flow of the algorithm that is utilised in the random forest technique can be seen in Figure.

### E. SVM and Naive Bayes

The classification and regression tasks are best handled by the supervised learning method known as the support vector machine (SVM). Although SVM can also be used for regression analysis, its primary application is classification. The data is being partitioned into categories by the hyperplane that exists in the n-dimensional space. The hyperplane has two edges on either side. The accuracy of the classification is directly correlated to the size of the margin that surrounds the text. The model's goal is to extend the margin's width, so it can accommodate more data. The characteristics that are included in the dataset will determine the amount of dimensions that the hyperplane will have. The hyperplane appears as a dot in a space with only one dimension. It takes the form of a line in two-dimensional space and a plane in three-dimensional space. The SVM classifier model is applied to the UNSW-NB15 dataset as well as the NSL-KDD dataset in the work that is being suggested by Gu, J., et all., 2021 [5]. SVM is applied to the data once it has been converted in order to efficiently detect intrusion, and the accuracy of the detection is tested and proven [14].

The well-known classification method known as Naive Bayes is based on the Bayes theorem and operates in accordance with its principles. It is a model that is based on probability, and its purpose is to determine the likelihood of y given all of the X. Despite the fact that the model is most commonly employed for classification, it will be applied here for feature transformation. In this case, the naive Bayes technique is employed for feature transformation in order to improve the data quality prior to performing intrusion detection using the support vector machine (SVM). Before beginning the process of intrusion detection, it is necessary to engage in the activity of feature transformation as a type of pre-processing in order to improve the overall quality of the data.

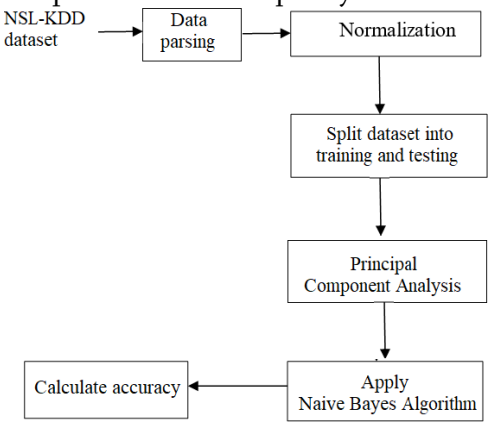


Fig 2.5: Naive Bayes based IDS

**F. Intrusion Detection Using Naive Bayes Modification**

The naive Bayes classification algorithm is a well-known classification approach that performs well in accuracy and other related measures. In addition, this algorithm is quite straightforward to implement. There are problems with the method whenever the probability is 0 with a certain variable. The problem is solved in Kurniawan, Y. I., et all., 2021[6] by disregarding the variable that is creating the zero probability and by amending the base theorem formula in order to make the naive Bayes algorithm suitable for the zero probability situation. Both of these solutions may be found in the article. If and only if the model experiences zero probability error based on the data, it will ignore the variable or make a change to the formula. The method's effectiveness in detecting intrusions has been demonstrated, along with its accuracy, recall, and precision, which have all been computed. Rootkit, Teardrop, Smurf, Neptune, and Satan are some of the attacks that can be detected by an intrusion detection system that uses a modified version of the naive Bayes method. The following table provides a comparison of the benefits and drawbacks of several algorithms that can be utilised for developing intrusion detection by making use of machine learning techniques.

Table 2.1 Comparison of different Algorithms

| Algorithm used | Paper | Advantages | Disadvantages |
|---|---|---|---|
| Logistic Regression -based Intrusion Detection using Fuzzy C-Means | P.kanimozhi et all.,2021. | Easy to use; effective in training | highly susceptible to overfitting when there are very few tuples |
| Intrusion Detection using KNN | Pan, J. S., et all.,2021 | It is easy to use and intuitive. | Dimensionality's curse |
| Intrusion detection using decision trees | Panigrahi , R., et all., 202 | Because scaling and normalisation are not necessary, preprocessing is simple. | extremely susceptible to changes in training data |
| Intrusion detection system using random forest algorithm | Chen, Z., et all., (2021) | The model is resistant to overfitting. | The model operates slowly because of the large amount of trees in the forest. |
| **SVM and Naive Bayes** | Gu, J., et all., 2021 | SVM operates better in large spaces with high dimensions. | Performance of SVM is not improved while dealing with huge datasets. |
| **Intrusion Detection Using Naive Bayes Modification** | Kurniawan, Y. I., et all., 2021 | Even when there is little data, naive bayes performs better. | Naive Bayes is always associated with the zero frequency problem. |

## III. CONCLUSION

This study provides an analysis of the various categorization approaches based on machine learning that can be used to identify intrusion in network traffic. The following classification algorithms are acceptable for use in intrusion detection using machine learning: logistic regression, K-nearest neighbour, decision tree, random forest, and support vector machine (SVM). The aforementioned methodologies and the roles they play in classification and the detection of intrusions have been thoroughly researched and examined from a variety of perspectives. The many study articles examine and show both the positive and negative aspects of the methodologies that are employed in the research. The KNN algorithm is straightforward, but when the

number of dimensions of the data is expanded, it does not perform particularly well. Logistic regression, despite its ease of implementation and widespread use, poses a significant risk of overfitting when applied to intrusion detection applications. The decision tree approach requires less pre-processing work, but it is susceptible to errors caused by variations in the data used for training. Random forest is not significantly impacted by overfitting in any way. When working with high-dimensional data, SVM always works quite well. The article does a very good job of discussing and analysing the work that solves the zero-frequency problem associated with naive Bayes.

References


[1] KANIMOZHI, P., ARULDOSS ALBERT VICTOIRE, T. (2021). OPPOSITIONAL TUNICATE FUZZY C-MEANS ALGORITHM AND LOGISTIC REGRESSION FOR INTRUSION DETECTION ON CLOUD. CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE, E6624.

[2] PAN, J. S., FAN, F., CHU, S. C., ZHAO, H. Q., & LIU, G. Y. (2021). A LIGHTWEIGHT INTELLIGENT INTRUSION DETECTION MODEL FOR WIRELESS SENSOR NETWORKS. SECURITY AND COMMUNICATION NETWORKS, 2021.

[3] PANIGRAHI, R., BORAH, S., BHOI, A. K., IJAZ, M. F., PRAMANIK, M., KUMAR, Y., & JHAVERI, R. H. (2021). A CONSOLIDATED DECISION TREE-BASED INTRUSION DETECTION SYSTEM FOR BINARY AND MULTICLASS IMBALANCED DATASETS. MATHEMATICS, 9(7), 751.

[4] CHEN, Z., ZHOU, L., & YU, W. (2021). ADASYN-RANDOM FOREST BASED INTRUSION DETECTION MODEL. ARXIV PREPRINT ARXIV:2105.04301.

[5] GU, J., & LU, S. (2021). AN EFFECTIVE INTRUSION DETECTION APPROACH USING SVM WITH NAÏVE BAYES FEATURE EMBEDDING. COMPUTERS & SECURITY, 103, 102158.

[6] KURNIAWAN, Y. I., RAZI, F., NOFIYATI, N., WIJAYANTO, B., & HIDAYAT, M. L. (2021). NAIVE BAYES MODIFICATION FOR INTRUSION DETECTION SYSTEM CLASSIFICATION WITH ZERO PROBABILITY. BULLETIN OF ELECTRICAL ENGINEERING AND INFORMATICS, 10(5), 27512758.

[7] KATHIRESAN,V, KARTHIK, S. (2021). EFFICIENT DETECTION USING SOFT COMPUTING APPROACH OF MODIFIED FUZZY C-MEANS BASED OUTLIER DETECTION IN ELECTRONICS PATIENT RECORDS SYSTEMS. JOURNAL OF MEDICAL IMAGING AND HEALTH INFORMATICS, 11(10), 2660-2666.

[8] DENNING, D. E. (1987). AN INTRUSION- DETECTION MODEL. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, (2), 222-232.

[9] ZHANG, X., LI, C., ZHENG, W. (2004, SEPTEMBER). INTRUSION PREVENTION SYSTEM DESIGN. IN THE FOURTH INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION TECHNOLOGY, 2004. CIT'04. (PP. 386390). IEEE.

[10] TSAI, C. F., HSU, Y. F., LIN, C. Y., & LIN, W. Y. (2009). INTRUSION DETECTION BY MACHINE LEARNING: A REVIEW. EXPERT SYSTEMS WITH APPLICATIONS, 36(10), 11994-2000.

[11] DANIELSSON, P. E. (1980). EUCLIDEAN DISTANCE MAPPING. COMPUTER GRAPHICS AND IMAGE PROCESSING, 14(3), 227-248.

[12] P. DIVYA, D. PALANIVEL RAJAN, N. SELVA KUMAR . (2021). ANALYSIS OF MACHINE AND DEEP LEARNING APPROACHES FOR CREDIT CARD FRAUD DETECTION. IN ICCCE 2020 (PP. 243-254). SPRINGER, HTTPS://DOI.ORG/10.1007/978-981-15-7961-5_24.

[13] SOMEFUN, O. A., AKINGBADE, K., & DAHUNSI, F. (2020). THE LOGISTIC-SIGMOID FUNCTION. ARXIV EPRINTS, ARXIV-2008.

[14] Noble, W. S. (2006). What is a support vector machine? Nature biotechnology, 24(12), 1565-1567.

[15] Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012, July). How many trees in a random forest? In International workshop on machine arning and data mining in pattern recognition (pp. 154-168). Springer, Berlin, Heidelber