Journal of Theoretical and Applied Information Technology
31st August 2022. Vol.100. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GRADIENT BOOSTED DECISION TREE (GBDT) AND GREY WOLF OPTIMIZATION (GWO)BASED INTRUSION DETECTION MODEL

MADURI MADHAVI [1], DR. NETHRAVATHI[2]

Research scholar, Reva University, Department of CSE,India
Professor, Reva University, Department of CSE,India
Email: [1] madurimadhavi@gmail.com, [2] nethravathi.np@reva.edu.in

ABSTRACT

Intrusion Detection Systems (IDS) have become increasingly important as computer networks have grown in size and complexity over the last several years, and they have become an essential component of the system's infrastructure. Increases in the number of breaches into computer networks have been attributed to a thriving underground cybercrime market and advanced technologies that make it simpler to hack into computer networks over the past decade. Researchers in both industry and academia have been working on strategies to detect and prevent these types of security breaches for more than 40 years, and the results have been promising. They have also put in place systems to assist them in this endeavour. It is critical for an intrusion detection system (IDS) to be able to deal with situations such as a low detection rate and a large amount of work. One of the most pressing concerns facing the globe today is that of data security. The hacking of data can occur in a variety of ways, which can render any network or system less effective. Interceptor denial of service (IDS) systems that are still in use are incapable of keeping up with the continually changing and complex nature of incursion activities on computer networks. Discovering and preventing these types of attacks is one of the most difficult tasks we face today. Machine Learning has risen to become one of the most successful methods of learning about things in recent years, thanks to advances in artificial intelligence. Machine Learning approaches have the ability to generalise from known attacks to variations, or even discover new sorts of breaches, which is a beneficial thing for security professionals. In recent years, there has been a lot of research on how to combine multiple strategies in order to increase the detection rates of Machine Learning classifiers. Since the development process for intrusion detection systems has included the use of Artificial Neural Networks and Decision Trees, the technology has risen in popularity. In order to determine how well different machine learning classifiers performed with the KDD99 incursion dataset, a number of tests and evaluations were conducted. The study's main focus was on Grey Wolf optimization (GWO) models based on Gradient Boosted Decision Trees (GBDT). They were used to identify and categorise intrusions, among other things. Furthermore, this research looked into how to create attack rules by using a KDD99 dataset to look for anomalies in network audit data, which was then used to create attack rules. Performance indicators such as the false negative and false positive detection rates are important in order to increase the rate at which an intrusion detection system detects objects that have been broken into by an intruder. In the experiments, different algorithms are compared. There is evidence to suggest that the GBDT is the best since it is the most accurate and does not produce many false positives.

Keywords: *Gradient Boost, Decision Tree, Intrusion Detection System, Grey Wolf Optimization, Machine Learning*

1. INTRODUCTION

The rapid growth of the Internet has resulted in a wide range of network security issues. In addition, a variety of security measures have been put in place to keep computer systems safe. Intrusion detection systems are one of these tools. Defending computer networks against intruders and attackers is a primary responsibility of intrusion detection systems (IDSs).An IDS's primary function is to detect and prevent abnormal network traffic connections. One of the most common approaches to detecting intrusions is to look for misuse, while the other is to look for anomalies in the data. By matching a network connection's characteristics to known intrusion signatures, a misuse-based detection method can identify an attack before it has occurred. If this connection matches one signature, it is an attack. With a low false alarm rate, these techniques are useful in identifying known network threats. But

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

when confronted with a new type of intrusion that has properties that don't match any of those in the database, they will fail.

A misuse-based method must update its database on a regular basis to address this issue. AS a result of this, obtaining the signatures of newly discovered intrusive behaviours can be extremely costly. On the other hand, a detection based on anomalies the method trains models of normal behaviour and identifies anomalies. Spotting suspicious traffic patterns in the network deviation from the standard profile the underlying assumption in Detection methods based on anomalies are a characteristic of an When compared to normal connections, abnormal connections are worlds apart. Anomaly detection is no longer necessary in an anomaly-based system. Signatures of attack patterns are needed to keep track of. not only unknown anomalies can be detected by anomaly-based detection methods. Unwelcome habits, but there's also no telling what might happen in the future. That's what I mean. Over misuse-based approaches Despite this, they continued to do so.it's possible to misclassify some of the more typical through traffic in the normal and abnormal behaviour boundaries. Because Anomaly-based techniques are becoming more common as new threats emerge. Drew more research attention. Great efforts have been made to achieve reliable detection results. Rule-based systems are best in the early stages of development. Statistical analysis and expert systems were both utilised. When dealing with large-scale network traffic, however, these approaches are only useful for small datasets. Intrusion detection based on anomaly is a classification issue in nature.

As a result, academics have built IDS systems that incorporate a variety of machine learning methodologies, including as decision trees, support vector machines (SVMs), naive Bayes, and artificial neural networks, among other techniques (ANNs). The major purpose of an IDS is to be as precise as possible while still remaining cost-effective in its design. Numerous attempts have been made, with varying degrees of success, to increase the precision of intrusion detection systems by including a variety of machine learning algorithms into the system. With a hybrid model, one of the primary goals is to bring together several different machine learning algorithms in order to dramatically increase detection performance. These techniques, on the other hand, are only appropriate for small datasets when dealing with large amounts of network traffic. They are not suitable for large datasets. In the natural world, intrusion detection is a classification difficulty that must be overcome. In order to get the highest level of accuracy possible when developing an IDS, the primary goal must be achieved. So hybrid approaches have become increasingly popular as a means of improving intrusion detection accuracy as compared to the use of separate machine learning methodologies.

Hybrid models are used to demonstrate a novel approach to network intrusion detection. In this research, the GBDT-GWO hybrid model is proposed as an unique hybrid model for solving IDS with high accuracy. One of the algorithms employed in the suggested method is GBD (gradient-booted decision trees), which is a type of decision tree. It is often referred to as GBRT (gradient-booted regression tree) or MART (multi-objective regression tree) (multiple additive regression tree). This approach makes use of the GWO algorithm (Grey Wolf optimization (GWO) as well as the GWO algorithm. It is possible to shorten the learning time of an intrusion detection classifier by training it on the optimal subset of features, while simultaneously enhancing accuracy. As a result, feature selection is an extremely important step in the process. The suggested method, which extracts key features from source datasets using a dimensionality reduction technique known as GBDT, makes use of the GBDT technique. Following the selection of features in the manner described above, the GBDT technique is used to train a prediction model on the optimal feature space identified. Training the prediction model on the ideal feature space is accomplished through the use of the GBDT methodology. The GBDT technique, which is an incredibly successful supervised learning method, must be utilised in order to combine the gradient boosting framework and the decision tree methodology into a single ensemble model that incorporates both tactics. GBDT algorithms are used in a variety of applications, including disease modelling, web search ranking [29, 30], and trip time prediction. We anticipate that GBDT algorithms will be an ideal fit for intrusion detection. Several GBDT parameters, notably the learning rate, are optimised using the GWO algorithm in order to achieve optimal efficiency. The NSL-KDD dataset has been used to test this methodology, which is compared to a variety of different baselines, including single classification methods

and hybrid models, in order to assess how well it performs in terms of accuracy. According to the experimental results, the suggested hybrid technique outperforms the baseline detection methods in terms of detection accuracy.

2. LITERATURE SURVEY

In recent years, researchers have devoted particular attention to the methodology used to choose features for network intrusion detection systems (IDS). For boosting network IDS performance, the researchers offered several ways that included various filtering and wrapping techniques as well as data processing and optimization methods as well as mechanism knowledge methods and bio-inspired metaheuristic procedures. It is possible to boost the performance of network intrusion detection systems by including bio-inspired Metaheuristic algorithms, which are well-known for their ability to discover the most effective solutions in the shortest amount of time. Each bio-inspired metaheuristic algorithm has a unique set of drawbacks and benefits that set it apart from the others. Hybridization allows each algorithm to benefit from the strengths of the others while also addressing the deficiencies of each algorithm individually. Much recent research has revealed that combining bio-inspired metaheuristic algorithms with other algorithms improves their overall performance. Several of this new research are discussed in greater detail in this section.

Kim (2014) and colleagues developed a hybrid intrusion detection system (IDS) that detects intrusions by utilising a multi-class SVM-based anomaly detection algorithm. The C4.5 algorithm-based decision tree model for detecting abuse is a decision tree model that is based on the C4.5 algorithm. It is used to detect misuse. The proposed hierarchical model was validated using the NSL-KDD dataset to determine its detection accuracy and false alarm rate for both unknown and known assaults, as well as for known assaults in general. When compared to existing models, the suggested model has the potential to significantly reduce the number of false positives while simultaneously decreasing the amount of testing and training time necessary. Additionally, as compared to present practises, the proposed technique significantly reduces the time required for training procedures by 50% and the time required for testing processes by 40%.Eesa et al. (2015) built a hybrid model by combining the cuttlefish optimization strategy (CFA) with a decision tree classifier in order to achieve higher classification accuracy in their research.Its goal is to spot network breaches. The decision tree method was used to identify the categories of aberrant occurrences in this model, while the CFA was utilised to select relevant features. On the KDDCup99 Dataset, the model's performance was evaluated. The findings revealed that the detection rate and accuracy are much higher when the number of characteristics is less than 20.

Irregularities in large-scale datasets are categorised using an IDS hybrid developed by Ghanem (2015) and colleagues. The hybrid integrates Genetic Algorithm (GA) detectors with a multi-start metaheuristic system to find and categorise irregularities. The proposed model makes use of a technique for creating detectors that is based on negative selection. The dataset used for this evaluation was the NSL-KDD dataset. It appears that the model is effective in terms of supplying an acceptable number of detectors, based on the evaluation results. This model has a 96.1 percent accuracy rate and a 3.3 percent false positive rate. A hybrid model for recognising intrusions was developed by Guo (2016) and colleagues, which merged the capabilities of misuse-based and anomaly-based detection approaches to provide a more comprehensive detection solution. In this paradigm, a total of two irregularity uncovering mechanisms and one misuse detection component are used to detect anomalies (MDC). Broken connections are identified and repaired by ADC one using the ADBCC technique, which is developed by ADC one. The links that were determined to be aberrant and normal were then transmitted to the ADC two and the MDC for evaluation using the K-NN method, which was performed simultaneously. With the help of the KDDCup99 and the Kyoto University Benchmark Dataset, researchers were able to determine the efficiency of this hybrid method (KUBD). A recent experiment done utilising the KDD99 dataset revealed that the proposed model is capable of identifying both unknown and recognised threats with excellent accuracy and efficiency. Because of its high detection accuracy and low false positive rate, it is capable of detecting network anomalies in a timely and effective manner, resulting in increased productivity. This was demonstrated in an experiment conducted using the KUBD dataset, which demonstrated that the proposed model was significantly more effective in recognising attack traffic that lacked a specific label when compared

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to KDDcup99 and KDD99, respectively, in recognising attack traffic that lacked a specific label.Using Asahi-Shahri et al. (2016) as an example, they constructed a hybrid model that incorporated GA and SVM methodologies to accomplish their results. The number of attributes in this model has been decreased from 45 to ten, a significant reduction. After separating the features into three groups depending on their importance, the GA algorithm determined that the first category was the least significant of the three. As demonstrated by its high true positive rate and low false positive rate when applied to the KDD 99 dataset, this model is effective at detecting false positives. According to the results of the proposed hybrid model, the true positive value was 0.973 and the false positive value was 0.017. This was revealed after the model was applied.

They released a model in 2018 that takes the Gini index into consideration, which was established by Li and colleagues. To deal with the problem, a gradient boosting decision tree (GBDT) and a PSO are used in cooperation with one another to find a solution. This index was used to assess which qualities should be included in the final product and which aspects should be omitted from it, based on the data. It was necessary to detect a network attack in order to prevent it from occurring. The gradient lifting decision tree technique was employed to accomplish this. The PSO technique was used for the goal of optimising the GBDT's settings to achieve the best results. Among other things, we looked at the model's detection rate, accuracy, F1-score, precision, and false alarm rate, among other things. This evaluation was carried out using the NSL-KDD Dataset, which can be found here. They released a model in 2018 that takes the Gini index into consideration, which was established by Li and colleagues. To deal with the problem, a gradient boosting decision tree (GBDT) and a PSO are used in cooperation with one another to find a solution. This index was used to assess which qualities should be included in the final product and which aspects should be omitted from it, based on the data. It was necessary to detect a network attack in order to prevent it from occurring. The gradient lifting decision tree technique was employed to accomplish this.

The PSO technique was used for the goal of optimising the GBDT's settings in order to achieve the best results. Among other things, we looked at the model's detection rate, accuracy, F1-score, precision, and false alarm rate, among other things. This evaluation was carried out using the NSL-KDD Dataset, which can be found here.. The findings demonstrated that the model is accurate and capable of appropriately identifying intrusions in a network environment. 78.48 percent of the time, the model detected the threat, and 96.44 percent of the time, it had an F1-score of 86.54 percent, and it had a rate of 3.83 percent false acceptance.

Several researchers, including Khraisat et al. (2020), have created a hybrid illness detection system (HIDS) that combines a C5.0 decision tree classifier with a one-class support vector machine to detect a wide spectrum of illnesses (OC-SVM). A hybrid intrusion detection system (HIDS) is a combination of intrusion detection systems that are based on signatures and intrusion detection systems that detect anomalies. While the C5.0 decision tree classifier was used to produce the signature-based intrusion detection system, it was also utilised to develop the anomaly-based intrusion detection system, which was created using the OC-SVM algorithm (object-oriented support vector machine). It is necessary to have a high detection rate while also having a low number of false alarms in order to detect known intrusions as well as zero-day attacks. It is also necessary to maintain a low rate of false alarms in order to attain this objective. HIDS datasets from the NSL-KDD as well as the Australian Defence Force Academy (ADFA) were used in the evaluation of the proposed system.

According to the results, HIDS outperforms both signature-based and anomaly-based intrusion detection systems in terms of the detection rate, false alarm rate, true negative rate, false-negative rate, false positive rate, and recall rate. It also outperforms them in terms of precision, sensitivity, and F-measure.A hybrid classification model was developed by Hajisalem (2018) and colleagues, which integrates approaches from the artificial bee colony (ABC) and the artificial fish swarm (AFS) studies, among others (AFS). Two datasets were used to evaluate the model's performance for the goal of determining its overall effectiveness (NSL-KDD and UNSW-NB15). During the trials, it was demonstrated that the model had a detection accuracy of about 100 percent, which was supported by the research data. False positives occur at a rate of less than 0.01 percent, which is extremely low and is considered exceptional.

Following the suggestion and development of a range of hybrid intrusion detection systems by Srivastava and colleagues, which are discussed in detail below (2019). To improve the accuracy of classification, several algorithms, including the entropy basic graph (EBG), the support vector machine (SVM), the generalised regression neural network (GRNN), and the k-nearest neighbour method (k-NN), were combined with the grey wolf optimization (GWO) method. The result was a more accurate classification than before (KNN). It was decided to use the KDD-99 dataset to investigate the classification of data into two categories: normal and intrusive. Normal data was classified into two categories: The data was analysed using a number of different categorization techniques. Aside from that, they rift the taxing data into numerous capacities and examine how well the proposed method performs in each volume of data. The findings indicate that, when compared to other classification approaches, the GWO–EBG classification methodology yields the best accurate results in this investigation, which is consistent with previous findings.

3. PROPOSED GBDT-GWO MODEL

The suggested method's key goals are: 1) The detection of insensible malicious actions; 2) The detection of malicious activities without the need for a deep packet inspection; and 3) The main features of the proposed detection strategy are shown in Figure 1.The Gradient Boosting Decision Tree (GBDT) was a big hit in a lot of places. In terms of the GBDT, it is a decision tree ensemble model. GBDT learns the decision trees throughout each iteration. Fitting the negative gradients accomplishes this. Despite this, as the amount of data grows, GBDT's efficiency and accuracy confront issues and hurdles. The computing difficulties of GBDT, for example, are proportional to the number of instances and features. As a result, multiple time-consuming calculations are required.The Light approach was offered as a solution to these problems. This approach is a distributed framework for boosting the gradient that is built on a decision tree for the GBM implementation.

In compared to other GBMs, LightGBM's approach has undergone some refinement. It is a decision-making algorithm that is based on a histogram-based decision tree and uses histogram subtraction for acceleration. Through the application of the Leaf-wise leaf development technique with depth limitation, it led to the optimization of sparse features. As a result, the number of errors will be reduced. It will increase the level of precision. In terms of the Leaf-wise depth limitation, it can ensure a high-efficiency level. At the same time, it can prevent over-fitting. The cache hit rate was improved, and the multi-threading was improved.

3.1 Dataset Preparation

In machine learning and data analysis techniques, this stage is very crucial. Preparing data entails transforming and presenting information in a suitable format. Particularly when the data is in a variety of formats and has a large range of information values. The following steps make up this level.

3.1.1 Data Transformation

Among other things, the KDD99 dataset contains a vast number of features and data that are presented in a variety of formats. These formats include script, numbers, signs, and other symbols, among others. These features' examination could take a long time to process and demand a lot of hardware resources. As a result, the transformation procedure was created to convert symbolic features to numeric features to avoid these snags.

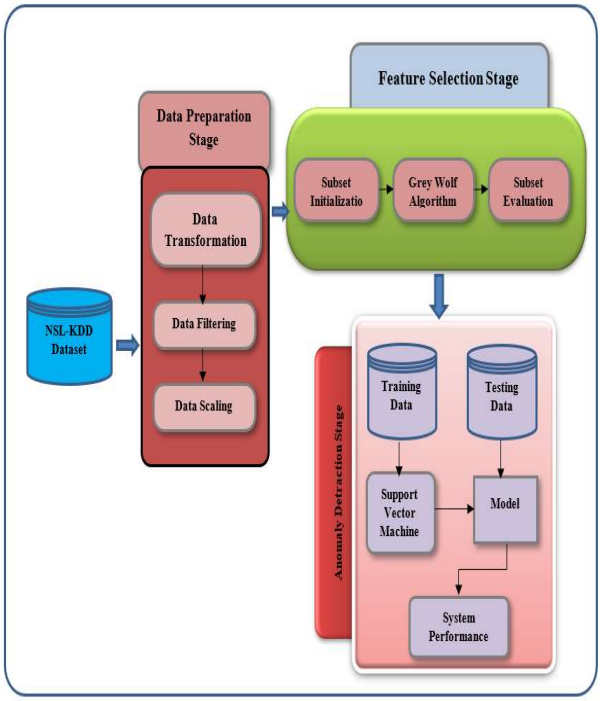

Fig 1.

Proposed system GBDT-GWO

3.1.2 Data normalisation

According to the definition, this approach is a way for changing a wide range of feature values into a more balanced range of values. As part of this investigation, Eq. is used to scale each value in the feature record (1).

$$Y^{-} = \frac{Y}{Y_{-maximum}}$$

As stated in the definition, this method involves turning an extremely wide range of feature values into a more balanced range of feature values. As part of this investigation, the equation is used to scale each value in the feature record (1).

3.1.3 Data Filtering

Data is typically selected or deleted from a dataset through the use of the filtering process. It was decided to use the filtering method to abstract and separate various sorts of spells in order to validate the suggested methodology in a variety of attack situations in this work. It is possible to find a large variety of assault kinds in the NSL–KDD dataset. Additionally, each sub-attack is associated with a certain type of dataset assault, such as a Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R) attack, among others (U2R). This method resulted in the assignment of each attack class to a principal attack category, which was then assigned to each attack class. Finally, the output contains a range of dataset containers, each of which has been subjected to a different type of dataset attack. This concludes the analysis.

3.2 Feature Selection

When a data set has been prepared, the feature collection stage is used to select the most valuable subsection of characteristics from among a wider range of characteristics. In this effort, the GWO method was updated in order to identify the best collection of structures possible. Detailed information on the stage 2 procedures can be found in the subsections that follow this one.

3.2.1 Subset Generation

There are several types of heuristic search strategies, but the most common is subset generation, which uses samples from the search region to generate candidate solutions for the subset evaluation. This method is used to identify answers for a range of problems. After much consideration, it was decided to use the accidental subgroup generation method (Kim et al. 2010) in order to build a subset of attributes.

$$Y_{(i,j)}(1) = y_j^{\min} + \delta(y_j^{\max} - y_j^{\min}) \quad (2)$$

If you are creating the matrix during the initialization step, the dimension of the matrix is yij. The variables ymax j and ymin j denote the top and lower boundaries of the function, respectively. The input and output values are are integers with values ranging from 1 to N, respectively. (1–D). In this equation, N signifies the number of possible solutions, and D is the size of each solution in the matrix.

3.2.2 Grey Wolf Optimisation Algorithm

The GWO algorithm is a swarm-based algorithm that was proposed by a group of researchers [11]. It was the natural social behaviour of grey wolves that served as an inspiration for the GWO. The wolves' chasing and hunting behaviour in order to capture prey is a fantastic example of the pursuit of the greatest feasible choice in any situation. When grey wolves are in the wild, they like to live in groups called packs. Pack sizes range from five and twelve wolves on average, depending on the season. As an added bonus, wolves are divided into four groups based on where they are in the pack, which aids in the hunting process by restricting the number of wolves in each group, which is favourable during the hunting process . These organisations are referred to by the names listed below: This can be either a male or female Alpha. He or she is in charge of making choices regarding the pack's hunting, waking, and sleeping patterns as well as where the pack will hunt and how they will hunt. Additionally, beta is a second level of wolves made up of either male or female wolves that aid the other wolves in the pack in making decisions about their future. Delta is the third rank, and their responsibilities include caring for others, serving as sentinels, acting as pack elder, and hunting. The letter Omega ()

represents the pinnacle of the hierarchical structure. In the hierarchal paradigm, this is the weakest of the lves, and it serves as a scapegoat who must comply with the demands of others.

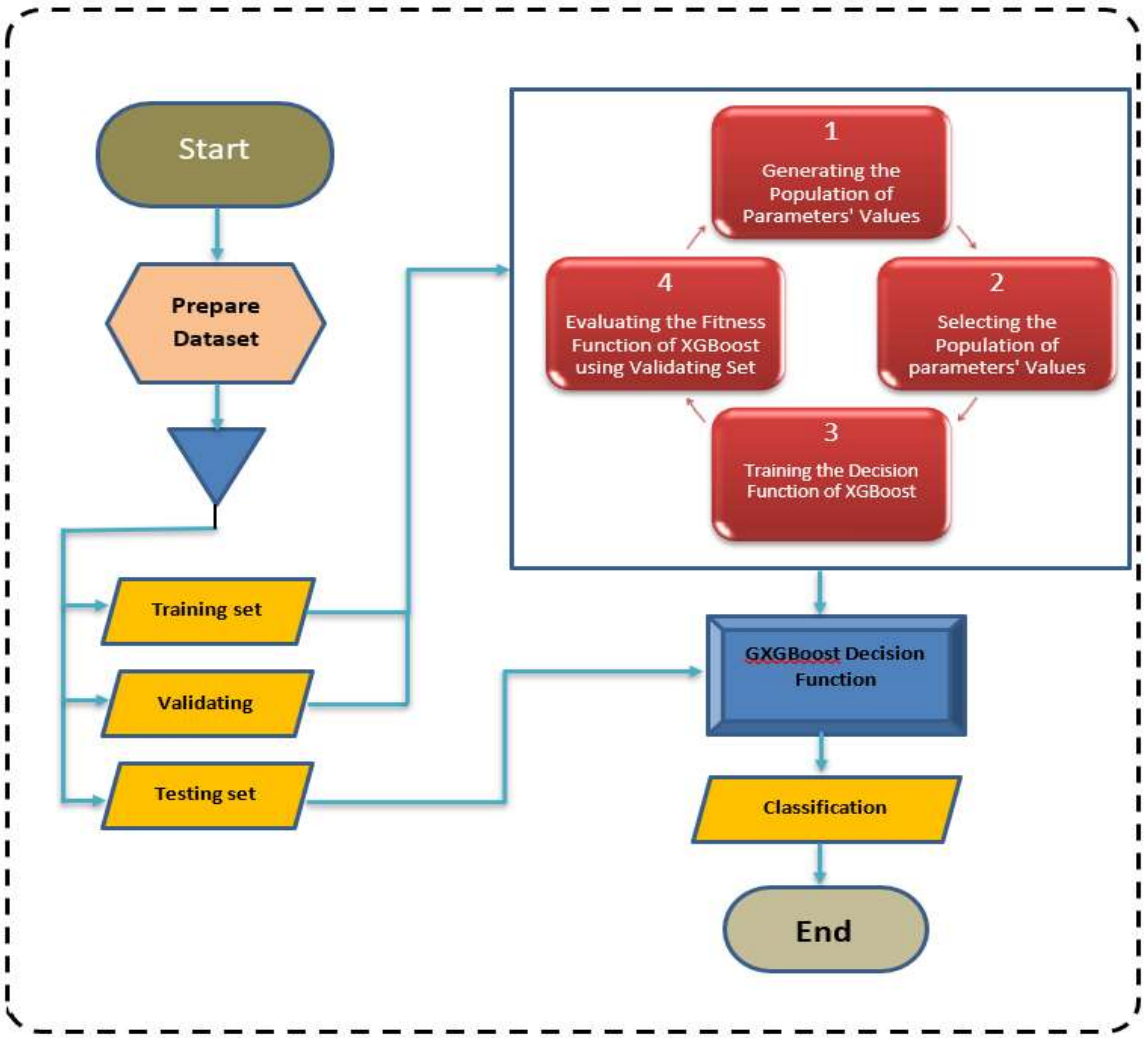

Fig.2 : Flow Model Of Proposed Systems

Level 3: Delta (δ): This word refers to wolves who are currently considerate of their fellow wolves and are not aggressive toward them (delta wolves). They keep a close eye on the x wolves, as well. The jobs they conduct range from scouting to sentinels to elders to in-pack caretakers to hunting and include a variety of tasks.

Level-4: It was believed that the wolves at the fourth degree of difficulty were the weakest. Omega (ω): They perform the function of scapegoat. They are required to follow the directions of specific individuals.

3.2.4 Mathematical Model And Algorithm

The GWO mathematics model is divided into three sections. Those components engage in behaviours such as surrounding, hunting, and adhering to other components. Encircling behaviour is represented by the equation on the right.

$$\overrightarrow{D} = \left|C.\overrightarrow{X_P}(t) - \vec{X}(t)\right| \quad (3)$$

$$\vec{X}(t+1) = \vec{X}(t) - \vec{A}.\vec{D} \quad (4)$$

where t is the current iteration, (A and C) denote the coefficient matrix vectors, Xp. denotes the prey position vector, and X denotes the grey wolf position vector The vectors will be discussed in greater detail in the following sections (A and C):

$$\vec{A} = 2\vec{d}.\overrightarrow{r_1} - \vec{d} \quad (5)$$
$$\vec{C} = 2\vec{d}.\vec{r_2}$$

Over the course of iterations, (a) reductions from 2 to 0, and the random vectors r1, r2 are generated. Secondly, there's the prey hunt: Grey wolves are capable of tracking down and hunting prey, and they frequently track the alpha wolf throughout the process. In addition, under specific conditions, the beta and delta may be permitted to participate in hunting on a limited basis. In general, alpha wolves are considered to be the most advantageous of the wolf pack, while beta and delta wolves are also thought to have considerable knowledge of prospective prey sites. Because of the placements of these three wolves, it will be possible to change the positions of the other wolves, including those of the omega wolf, as indicated in the following equations. In order to alter the locations of the other wolves, including the omega wolf, it will be necessary to employ the alpha, beta, and delta positions.

$$\vec{x}(t+1) = \frac{1}{3}\overrightarrow{X_1} + \frac{1}{3}\overrightarrow{X_2} + \frac{1}{3}\overrightarrow{X_3} \quad (7)$$

where X1, X2 and X3 are given by following Equations:

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha}(t) - \overrightarrow{A_1}.\overrightarrow{D_\alpha} \quad (8)$$
$$\overrightarrow{X_2} = \overrightarrow{X_\beta}(t) - \overrightarrow{A_2}.\overrightarrow{D_\beta} \quad (9)$$
$$\overrightarrow{X_3} = \overrightarrow{X_\delta}(t) - \overrightarrow{A_3}.\overrightarrow{D_\delta} \quad (10)$$

This iteration's placements of the alpha (first), beta (second), and delta (third) wolves, which correspond to the first three best solutions to our problem, are given by the letters (x), (y), and (z). Specifically, in Equations 8, 9, and 10, the variables A1, A2, and A3 are stated, respectively; and, in the following Equations, D 1, D 2, and D 3 are derived, respectively:

$$\overrightarrow{D_\alpha} = \left|\overrightarrow{C_1}.\overrightarrow{X_\alpha} - \vec{X}\right|$$

$$\overrightarrow{D_\beta} = \left|\overrightarrow{C_2}.\overrightarrow{X_\beta} - \vec{X}\right| \quad (12)$$

$$\overrightarrow{D_\delta} = \left|\overrightarrow{C_3}.\overrightarrow{X_\delta} - \vec{X}\right|$$

where C1, C2, and C3 are represented by the equations 11, 12 and 13. It demonstrates how the search wolves adjust their position in the search space when the values of alpha, beta, and delta change. Remember that the final position would be chosen at random from among the points on a circle formed by the alpha, beta, and delta coordinates of the search space, with the ultimate location being decided by the alpha, beta, and delta coordinates of the search space, which is critical to understand. The following is a concise summary of the situation: wolves such as the alpha, beta, and delta wolves select where prey is to be found, whereas the other wolves constantly alter their positions at random around the prey.

After the victim has stopped moving, the wolves proceed to attack it in a ferocious manner. In a given iteration, a decrease in value from 2 to 0 signifies that the wolves are getting closer to their meal. The following equation describes the value of an in terms of a:

$$\vec{a} = 2 - \frac{2*t}{Maxltr} \quad (14)$$

which specifies the current iteration, and MaxItr, which denotes the maximum number of iterations that can be performed simultaneously. When it came to feature selection, the GWO method was applied in this experiment. The results showed that it was a successful technique. Because of its behaviour based on meta-heuristics, it can find the best answer and avoid stacking on one option. Furthermore, it performs admirably in hitherto unexplored and difficult search locations. It also has a small number of control settings and is simple to implement. Furthermore, the GWO bases its selection on the top three search agents. This equation is used in the initialization step to change the answerproduced by GWO procedure to binary values, and it is also used to choice the most appropriate explanation in the finalisation stage, which is shown in the following equation.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$Z_i = round(|y_i \bmod 2|) \bmod 2 \quad (15)$$

When the binary value (discreate value) denoted by the numbers 0 or 1 is represented by Zi, I denotes the number of explanations, and yi denotes the worth of the explanation (continues values) formed during the initialization and final stages of the process, the process is said to be complete. From what we can tell thus far, Eq. (12) states that a binary number is zero if its absolute value falls between zero and 0.4999, or between 1.5 and 1.9999, and that a binary number is one if it falls between 1.5 and 1.9999. Furthermore, if the absolute value of the residual is between 0.5 and 1.49999, the binary number is 1, and if the absolute value of the residual is greater than 1.49999, the binary number is 0; furthermore, if the absolute value of the residual is greater than 1.49999, the binary number is 1. In the event that this is not the case, the value is set to zero.

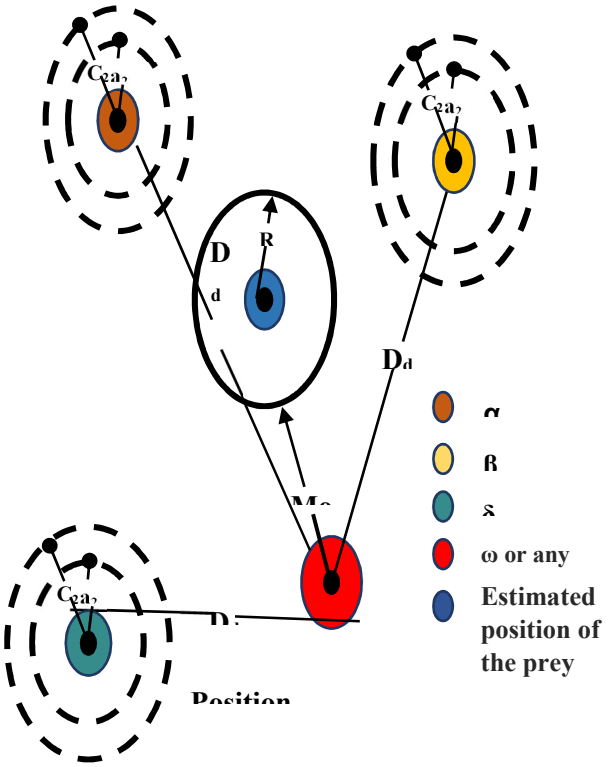

Fig 3.Iterative updates in position

x_, x_, and x_ are three letters that stand for x_, x_, and x_. The algorithm will also select t he first three best possibilities in terms of the placements of the alpha, beta, and delta wolves after one iteration, and the other wolves in the pack will pay attention to the first three best options in terms of their placements. Specifically, in the case of alpha's, first and foremost are the solutions (positions) with the highest classification accuracy, then beta's and finally delta's. Every time the process is performed, the classifier is trained and validated further to improve its performance. The accuracy of the classifier is then determined for each subset (solution) of the position matrix, and the process is repeated a second time for each subset (solution).

The Gradient Boosting Decision Tree that has been proposed features a fitness function that is weighted sum in nature (GBDT). In order to minimise the number of selected features while maintaining a high level of accuracy in categorization, the fitness evaluation in this study attempts to reduce the number of selected features while maintaining a high level of accuracy. In order to establish the fitness of a prospective solution, the accuracy and the quantity of features contained in the prospective solution are taken into consideration.

$$ProposedFitnessFunction = W_1 * Accuracy + W_2 * \frac{1}{Numberofselectedfeatures} \quad (16)$$

Each feature subset contains a list of the features that are currently accessible. A subset with the smallest number of characteristics is chosen when two subsets have identical accuracy but differ in the amount of attributes they contain. When (W1) is divided by Accuracy and (W2) is divided by the inverse Number of Selected Features, it is possible to customise the values of (W1) and (W2) in the preceding equation, as shown in the following equation.This has resulted in more weight being placed on accuracy (W1) rather than the number of selected features (W2) to be selected, due to the fact that accuracy (as opposed to the number of picked features) is the more critical issue among the two. Consequently, in all scenarios, (W1) should be greater than zero in all cases (W2).

4. RESULT AND DISCUSSION

4.1 Environment For Implementation

Data preparation, feature selection, clustering, classification, and other data analysis processes are carried out on a variety of platforms. Weka, knime, RapidMiner, MATLAB, and other programming tools are commonly cast-off in the machine learning field.The recommended approach is implemented in this work with the help of MATLAB. Moreover, 20 percent of the

KDD99 dataset was utilised to clarify the presentation of the suggested classical, and the tests were carried out on successive days on the

Windows 10 platform using a 3.2-GHz Core i7 processor with 12 GB of RAM, with findings published in the Journal of Machine Learning. The technique was further evaluated using a variety of different assault scenarios, and its performance was measured in terms of organization accurateness as fine as the amount of structures that were selected.

4.2 Assessment Criteria

Four criteria are computed to measure the effectiveness of the disease prediction process: accuracy, precision, recall, and f-score. The confusion matrix is commonly used in machine learning classifiers to derive these factors. The instances that make up the confusion matrix are divided into four groups:

- True Positive (TP) that represents aberrant cases that have been successfully classified.
- True Undesirable (TN) stands for "normal cases" that have been accurately classified.
- False Positive (FP) is defined as anomalous occurrences that have been categorized incorrectly.
- False Negative (FN) stands for "normal cases" that have been categorized incorrectly

4.3 Precision Analysis

This is accomplished by examining a range of performance metrics. A classifier's precision is defined by its accuracy across all classifications. Precision is a performance statistic that is used to evaluate the results of planned activities. This measure contrasts the overall number of TPs and TNs with the total number of correctly classified occurrences (True Negatives). The following equation illustrates the mathematical subject of precision.

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (17)$$

Here, the letters "TP" denote True Positives, whereas the letters "FP" denote False Positives.

Table 1. Precision analysis of GBDT-GWO technique with existing methods

Impact of Precision analysis in %					
No of data from datasets/ Methods	**Linear Regression**	**Random Forest**	**KNN**	**SVM**	**GBDT-GWO**
100	84.32	85.13	87.11	88.11	91.45
200	83.89	86.53	87.94	88.94	92.71
300	83.93	87.72	88.17	89.17	93.91
400	84.62	88.09	88.05	89.05	94.33
500	85.71	87.90	88.78	92.19	95.86

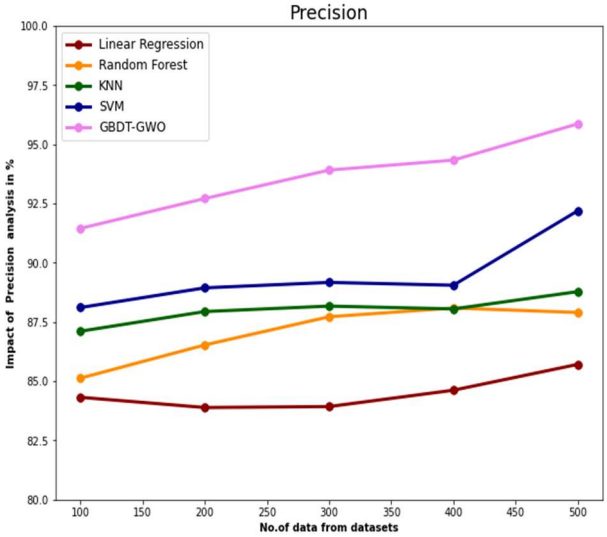

Fig. 4 illustrates a brief comparison of the GBDT-GWO technique to other existing techniques across a variety of datasets, demonstrating that precision is critical for improved outcomes. The figure indicated that the GBDT-GWO method performs best with the highest precision, whereas the Linear regression, Random forest, KNN and SVM techniques perform poorly. For example, the GBDT-GWO approach achieved a precision of 91.45% with 100 data from datasets, while the Linear regression, Random forest, KNN and SVM strategies achieved a precision of 84.32%, 85.13 %, 87.11 %, and 88.11%, respectively. Similarly, when 500 data are used, the GBDT-GWO technique achieves a precision increase of

95.86%, whereas the Linear regression, Random Forest, KNN and SVM strategies achieve a precision decrease of 85.71%, 87.90%, 88.78%, and 92.19%, respectively.

4.4 Recall Analysis

Recall is the next performance statistic that is determined for performance analysis. Recall is a metric that indicates the completeness of all classifiers. The mathematical equation is used to determine the value of Recall.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (18)$$

Where, FNdepicts the false negatives.

Table 2. Recall analysis of GBDT-GWO technique with existing methods

Impact of Recall analysis in %					
No of data from datasets/ Methods	**Linear Regression**	**Random Forest**	**KNN**	**SVM**	**GBDT-GWO**
100	84.32	85.13	88.21	89.21	91.35
200	86.19	87.53	88.92	89.91	91.82
300	86.83	87.72	89.17	90.11	92.73
400	87.52	88.79	90.25	90.95	94.61
500	88.23	89.90	91.17	93.64	96.62

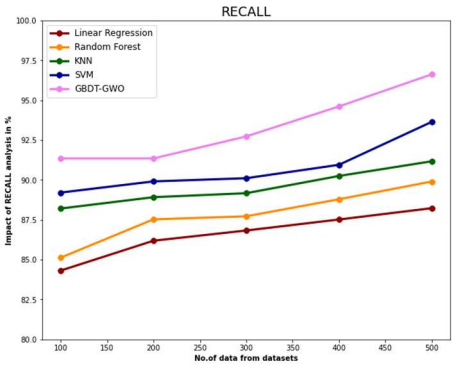

Fig 5. Recall Analysis Of GBDT-GWO Technique With Existing Methods

Table 2 and Fig. 5 illustrate a brief comparison of the GBDT-GWO technique to other existing techniques across a variety of datasets, demonstrating recall analysis for improved outcomes. The figure indicated that the GBDT-GWO method performs best with the highest recall, whereas the Linear regression, Random Forest, KNN and SVM techniques perform poorly. For example, the GBDT-GWO approach achieved a recall of 91.35% with 100 data from datasets, while the Linear regression, Random Forest, KNN and SVM strategies achieved a recall of 84.32%, 85.13 %, 88.21%, and 89.21 %, respectively. Similarly, when 500 data are used, the GBDT-GWO technique achieves a recall increase of 96.62%, whereas the Linear regression, Random Forest, KNN and SVM strategies achieve a recall decrease of 88.23%, 89.90%, 91.17%, and 93.64%, respectively.

4.5 F-Score Analysis

Overall, the F-score is a performance metric that considers the accuracy as well as the recall. It has a value ranging from 0 to 1 with a minimum value of 0. Calculated as the harmonic mean of the memory and accuracy numbers, it is a useful indicator of accuracy. The following is the formula for calculating the F-score:

$$\text{F}_{\text{score}} = 2.\frac{(\text{precision})(\text{recall})}{\text{precision} + \text{recall}} \quad (19)$$

Table 3. F-Score Analysis Of GBDT-GWO Technique With Existing Methods

Impact of F-Score analysis in %					
No of data from datasets/ Methods	Linear Regression	Random Forest	KNN	SVM	GBDT-GWO
100	84.12	87.21	88.18	89.72	91.25
200	85.89	86.85	85.89	88.96	91.82
300	84.23	87.48	86.21	89.26	93.04
400	85.92	88.27	86.95	91.78	95.81
500	87.95	89.55	90.32	91.76	95.38

Journal of Theoretical and Applied Information Technology
31st August 2022. Vol.100. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

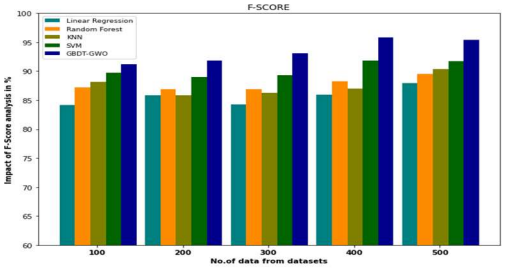

Fig 6. F-score analysis of GBDT-GWO technique with existing methods

Fig. 6and table 3 illustrate a brief comparison of the GBDT-GWO technique to other existing techniques across a variety of datasets, demonstrating the F-score for improved outcomes. The figure indicated that the GBDT-GWO method performs best with the highest F-score, whereas the Linear regression, Random Forest, KNN and SVM techniques perform poorly. For example, the GBDT-GWO approach achieved a F-score of 91.25% with 100 data from datasets, while the Linear regression, Random Forest, KNN and SVM strategies achieved a F-score of 84.12%, 87.21%, 88.18%, and 89.72%, respectively. Similarly, when 500 data are used, the GBDT-GWO technique achieves a F-score increase of 95.38%, whereas the Linear regression, Random Forest, KNN and SVM strategies achieve a F-score decrease of 87.95%, 89.55%, 90.32%, and 91.76%, respectively.

4.6 Accuracy Analysis

Accuracy is a measure of how close the suggested model is to the target value. In other words, it is a ratio that reflects the number of forecasts that have been produced in relation to the overall number of predictions. The correctness of the system is determined by applying the mathematical model revealed in the subsequent figure.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (20)$$

Where, TN denotes the true negatives

Table 4. Accuracy Analysis Of GBDT-GWO Technique With Existing Methods

Impact of Accuracy analysis in %					
No of data from datasets/ Methods	**Linear Regression**	**Random Forest**	**KNN**	**SVM**	**GBDT-GWO**
100	88.32	88.24	88.23	90.83	92.17
200	88.89	88.93	89.13	91.25	93.54
300	90.17	89.32	89.93	90.11	94.23
400	91.92	90.17	90.82	92.84	95.72
500	92.24	91.92	92.11	91.10	96.21

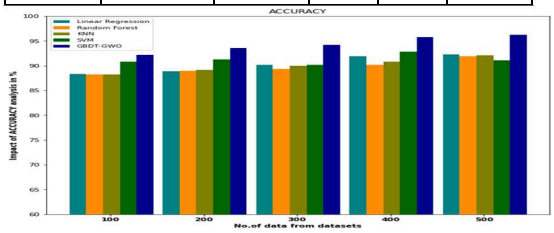

Fig 7. Accuracy Analysis Of GBDT-GWO Technique With Existing Methods

Table 4 and Fig. 7 illustrate a brief comparison of the GBDT-GWOtechnique to other existing techniques across a variety of datasets, demonstrating that accuracy should be a high priority for improved outcomes. The figure indicated that the GBDT-GWOmethod performs well with the highest accuracy, but the Linear regression, Random Forest, KNN and SVM techniques perform poorly. For example, the GBDT-GWOtechnique achieved an accuracy of 92.17 % with 100 datasets, whereas the Linear regression, Random Forest, KNN and SVM techniques achieved an accuracy of 88.32 %, 88.24%, 88.23%, and 90.83%, respectively. Similarly, when 500 data points were used, the GBDT-GWOtechnique increased accuracy to 96.21%, while the Linear regression, Random forest, KNN and SVM techniques decreased

accuracy to 92.24 %, 91.92%, 92.11%, and 91.10%, respectively.

4.7 Execution Time Analysis

The execution time (ET) is distinct as the period compulsory for a CPU to process instructions on a system. This is a critical metric for determining the proposed system's performance. This is the most precise numerical representation of the time complexity. The term TP (True Positive) refers to patients who are in good health and have been identified as such by the suggested procedure. The term TN (True Negative) refers to people who are in good health but have been classified as critical by the proposed method. Patients who were in critical health but were mistakenly recognized as normal by the proposed method were labelled as FP (False Positive). Finally, the term FN (False Negative) refers to patients who are in a severe health situation and have been identified as such by the proposed method.

Table 5. Execution time analysis of GBDT-GWO technique with existing methods

Impact of Execution Time analysis in ms					
No of data from datasets/ Methods	**Linear Regression**	**Random Forest**	**KNN**	**SVM**	**GBDT-GWO**
100	25.61	17.41	19.41	37.41	18.91
200	83.71	65.52	75.52	95.52	65.11
300	238.02	198.12	208.12	238.12	154.89
400	454.86	432.66	532.66	572.66	451.68
500	1112.56	1022.28	1162.2	1222.2	870.456

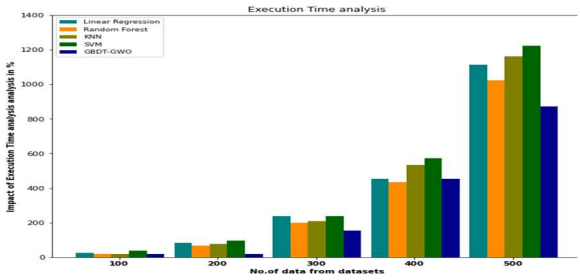

Fig 8. Execution Time Analysis Of GBDT-GWO Technique With Existing Methods

The GBDT-GWO technique is compared to various existing techniques in Fig. 8and table 5 while dealing with a different number of data points, and the execution time should be lowered for improved performance. The graph demonstrates that while the GBDT-GWO approach has the quickest execution time, the Linear regression, Random Forest, KNN and SVM algorithms produce a better-balanced outcome. For example, the GBDT-GWO technique processed 100 datasets in 18.91ms, whereas the Linear regression, Random Forest, KNN and SVM techniques took 25.61ms, 17.41ms, 19.41ms, and 37.41ms, respectively. Similarly, when 500 datasets are used, the GBDT-GWO technique takes only 870.456ms while the Linear regression, Random Forest, KNN and SVM procedures, with maximum execution times of 1112.56ms, 1022.28ms, 1162.2ms, and 1222.2ms, respectively.

5 CONCLUSIONS

The goal of this study is to assess whether or not a novel Gradient Boosted Decision Tree (GBDT) GWO method can be utilised to advance the recital of the IDS by including it into the decision tree. The Gradient Boosted Decision Tree GWO was used throughout the feature collection approach in order to provide an optimum subdivision of structures with high arrangement accurateness. This was recommended by the authors and was used throughout the feature selection procedure. The proposed technique was also tested on a 20 percent sample of the KDD99 dataset, which allowed for an evaluation of its overall performance. When conducting this study's analysis, the data separation technique was employed. When it comes to establishing the efficacy of an IDS organization and revealing its true presentation, difficult it to different sorts of network threats is a vital step. The proposed approach also revealed that it was capable of producing remarkable classification accuracy by utilising an ideal selection of features and a diverse range of assault circumstances. Additionally, the efficacy and feasibility of the proposed approach were proved through a comparison with modern approaches that yielded

superior outcomes. To better handle a wide range of optimization issues, the researchers are urged to combine the suggested Gradient Boosted Decision Tree with further bio-inspired algorithms in the future. Additionally, by including the most ideal GWO, these tactics might be used to recover the recital of the Gradient Boosted Decision Tree (GBDT) classifier, which would be a significant improvement. Additional Gradient Boosted Decision Tree features like as discovery rate, arrangement error, and so on will be incorporated into future study to further develop this topic further. Finally, we will make use of additional benchmark datasets that will include unique types of network assaults in their composition.

REFERENCES

[1] G. Kim, S. Lee and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," Expert Systems with Applications, vol. 41, no. 4, pp. 1690–1700, 2014.

[2] T. F. Ghanem, W. S. Elkilani and H. M. Abdul-Kader, "A hybrid approach for efficient anomaly detection using metaheuristic methods," Journal of Advanced Research, vol. 6, no. 4, pp. 609–619, 2015.

[3] B. M. Aslahi-Shahri, R. Rahmani, M. Chizari, A. Maralani, M. Eslami et al., "A hybrid method consisting of GA and SVM for intrusion detection system," Neural Computing and Applications, vol. 27, no. 6, pp. 1669–1676, 2016.

[4] W. L. Al-Yaseen, Z. A. Othman and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," Expert Systems with Applications, vol. 67, no. 4, pp. 296–303, 2017

[5] S. Hosseini and B. M. H. Zade, "New hybrid method for attack detection using combination of evolutionary algorithms, SVM, and ANN," Computer Networks, vol. 173, pp. 107–168, 2020.

[6] L. Li, Y. Yu, S. Bai, J. Cheng and X. Chen, "Towards effective network intrusion detection: A hybrid model integrating Gini index and GBDT with PSO," Journal of Sensors, vol. 2018, no. 6, pp. 1–9, 2018.

[7] V. Hajisalem and S. Babaie, "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection," Computer Networks, vol. 136, pp. 37–50, 2018.

[8] H. Mohmmadzadeh and F. S. Gharehchopogh, "A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: Case study Email spam detection," Preprints, pp. 1– 28, 2020.

[9] N. Moustafa and J. Slay, "The evaluation of network anomaly detection systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," Information Security Journal: A Global Perspective, vol. 25, no. 1–3, pp. 18–31, 2016.

[10] S. Mirjalili, S. M. Mirjalili and A. Hatamlou, "Multi-verse optimizer: A nature-inspired algorithm for global optimization," Neural Computing & Applications, vol. 27, no. 2, pp. 495–513, 2016.

[11] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," Knowledge-Based System, vol. 89, pp. 228–249, 2015.

[12] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," Computational Intelligence Magazine, IEEE, vol. 1, pp. 28-39, 2006

[13] Acharya N, Singh S (2018) An IWD-based feature selection method for intrusion detection system. Soft Comput 22:4407–4416

[14] Alamiedy TA, Anbar M, Al-Ani AK et al (2019) Review on feature selection algorithms for anomaly-based intrusion detection system.

[15] Kumar S, Joshi RC (2011) Design and implementation of IDS using snort, entropy and alert ranking system. In: 2011—international conference on signal processing, communication, computing and networking technologies, ICSCCN-2011. pp 264–268

[16] Wolf L, Shashua A (2005) Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weightedbased approach. J Mach Learn Res 6:378–384

[17] Vithalpura JS, Diwanji HM (2015) Analysis of ftness function in designing genetic algorithm based intrusion detection system. J Sci Res Dev 3:86–92

[18] Tribak H, Delgado-Márquez BL, Rojas P et al (2012) Statistical analysis of diferentartifcial intelligent techniques applied to intrusion detection system. In:

Proceedings of 2012 international conference on multimedia computing and systems, ICMCS 2012. pp 434–440
[19] Tavallaee M, Bagheri E, Lu W, Ghorbani AA (2009) A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE, pp 1–6
[20] Srivastava D, Singh R, Singh V et al (2019b) Analysis of diferent hybrid methods for intrusion detection system. 757–764
[21] Xingzhu W (2015) ACO and SVM selection feature weighting of network intrusion detection method. Int J Secur its Appl 9:259–2
[22] Xu H, Liu X, Su J (2017) An improved grey Wolf optimizer algorithm integrated with cuckoo search. In: Proceedings of the 2017 IEEE 9th international conference on intelligent data acquisition and advanced computing systems: technology and applications, IDAACS 2017. pp 490–493
[23] S. Eesa, Z. Orman and A. M. A. Brifcani, "A novel feature-selection approach based
[24] on the cuttlefish optimization algorithm for intrusion detection systems," Expert
[25] Systems with Applications, vol. 42, no. 5, pp. 2670–2679, 2015.
[26] Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman and A. Alazab, "Hybrid intrusion detection system based on the stacking ensemble of C5 decision tree classifier and oneclass support vector machine," Electronics, vol. 9, no. 1, pp. 173–191, 2020.
[27] J. Rene Beulah,L.Prathiba, G.L.N Murthy, E. Fantin Irudaya Raj and N. Arulkumar ," Blockchain with deep learning-enabled secure healthcare data transmission and diagnostic model", International Journal of Modeling, Simulation, and Scientific Computing, https://doi.org/10.1142/S1793962322410069
[28] Berlin, M.A., Tripathi, S. et al. IoT-based traffic prediction and traffic signal control system for smart city. Soft Computing (2021). https://doi.org/10.1007/s00500-021-05896-x
[29] Srivastava D, Singh R, Singh V (2019a) An intelligent gray wolf optimizer: a nature inspired technique in intrusion detection system (IDS). J Adv Robot 6:18–24
[30] Ozgür A, Erdem H (2017) The impact of using large training data set KDD99 on classification accuracy. PeerJPrepr5:e2838v1
[31] D Sivabalaselvamani, An automated learning model for sentiment analysis and data classification of Twitter data using balanced CA-SVM, *Concurrent Engineering Research and Applications*, Vol.29,No.4,pp 386-395.
[32] Bhukya, R. R., Hardas, B. M., Ch., T. et al. (2022). An Automated Word Embedding with Parameter Tuned Model for Web Crawling. Intelligent Automation & Soft Computing, 32(3), 1617–1632.
[33] Roopa Devi EM, Suganthe RC (2018) Enhanced transductive support vector machine classification with grey wolf optimizer cuckoo search optimization for intrusion detection system. ConcurrComput 1–11.
[34] A.V.R. Mayuri,Nilesh Shelke, An efficient low complexity compression based optimal homomorphic encryption for secure fiber optic communication, Optik, Vol 252,2022,pp.168545, https://doi.org/10.1016/j.ijleo.2021.168545
[35] D. K. Jain, S. K. K. Sah Tyagiand L. Natrayan, "Metaheuristic Optimization-based Resource Allocation Technique for Cybertwin-driven 6G on IoE Environment," in *IEEE Transactions on Industrial Informatics*, doi: 10.1109/TII.2021.3138915.
[36] Dastanpour A, Ibrahim S, Mashinchi R (2014) Using genetic algorithm to supporting artificial neural network for intrusion detection system. J.CommunComput 11:1–13
[37] Devi EMR, Suganthe RC (2017) Feature selection in intrusion detection grey wolf optimizer. Asian J Res Soc Sci Humanit 7:671