iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46940

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Intrusion Bit Detection in Data Packet Transmitted Over Covert channels using K-NN Machine Learning Algorithm

Raju Singh Kushwaha
Assistant Professor, Department of Computer Science and Engineering, Modern Institute of Technology and Research Centre, Alwar, Rajasthan, India

Abstract: ***K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning techniques. The algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well good category by using an algorithm. This algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems. If the encrypted data is transferred over the covert channel, the number of intruders, cryptanalysis attack and want to know data pattern by analysing the data bit pattern. If the intruder inserts any bit inside the data bit transferred to the receiver, then the receiver encryption system detects this bit and neglects this data for the decryption process.***
Keywords: KNN Algorithm, Intrusion, Covert Channel, Machine Learning.

I. INTRODUCTION

In the present era of the internet, the world of information is growing faster than anything else in the world. A variety of important data are transferred from sender to receiver. Sometimes the more important and secure data is transferred over the covert channel, this is a big threat to data confidentiality, authentication, and integrity.

To enhance the integrity, confidentiality, and authentication of data from intruders who trap the data bit and append some of the fake bit or distort the data for the receiver and also access some part of it and use it for some other unethical means.

This is also called the breach of security of data, which may harm any individual, organization, or institution by means of defamation, information, and impact of some economic loss.

The proposed algorithm KNN provides a solution by means of classification. The list of senders is already maintained in the system database and the machine learning algorithm firstly, identifies and classifies the sender, whether it is a real sender or a fake/spam sender, as per supervised learning of the algorithm, it accepts and then after that, it goes for next decryption process.

The decryption process involves the decryption of the message and its content based on the previously received message. The machine learning algorithm KNN classifies the pattern of learning from a previously received message from the sender and the current message and if any word is found to be not relevant to the context of the current message as well as the previously received message.

II. LITERATURE REVIEW

The problem of intruder data/bits in received messages is growing as per the growth of information networks. In this growing network of information, the security of data is also a concern to the sender, receiver as well as network service providers. There is some related work that applies machine learning methods to detect fake data.

1) They describe a focused literature survey of Artificial Intelligence Revised (AI) and Machine learning methods for email spam detection.
2) They have used the “image and textual dataset for e-mail spam detection with the employment of various methods.
3) They have used methods of KNN algorithm, Reverse DBSCAN algorithm with experimentation on a dataset. For text recognition, the OCR library” is employed but this OCR doesn't perform well.
4) They used the feature selection hybrid approach of TF-IDF (Term Frequency Inverse Document Frequency) and rough math.
5) The KNN algorithm description using a google search engine.

III. PROPOSED METHODOLOGY

A. *Intruder*

The most common threat to security is the attack by the intruder. Intruders are often referred to as hackers and are the most harmful factors contributing to the vulnerability of security. They have immense knowledge and an in-depth understanding of technology and security. Intruders breach the privacy of users and aim at stealing their confidential information of the users. The stolen information is then sold to third-party, which aim at misusing the information for their own personal or professional gains.

B. *Intruders Are Divided Into Three Categories*

1) *Masquerader:* The category of individuals that are not authorized to use the system but still exploit users' privacy and confidential information by possessing techniques that give them control over the system, such category of intruders is referred to as Masquerader. Masqueraders are outsiders and hence they don't have direct access to the system, their aim is to attack unethically to steal data/ information.
2) *Misfeasor:* The category of individuals that are authorized to use the system, but misuse the granted access and privilege. These are individuals that take undue advantage of the permissions and access given to them, such category of intruders is referred to as Misfeasor. Misfeasors are insiders and they have direct access to the system, which they aim to attack unethically for stealing data/ information.
3) *Clandestine User:* The category of individuals who have supervision/administrative control over the system and misuse the authoritative power given to them. The misconduct of power is often done by superlative authorities for financial gains, such a category of intruders is referred to as Clandestine Users. A Clandestine User can be any of the two, insiders or outsiders, and accordingly, they can have direct/ indirect access to the system, which they aim to attack unethically by stealing data/ information.
4) *Brute Force Attack:* A brute force attack is a hacking method that uses trial and error to crack passwords, login credentials, and encryption keys. It is a simple yet reliable tactic for gaining unauthorized access to individual accounts and organizations' systems and networks. The hacker tries multiple usernames and passwords, often using a computer to test a wide range of combinations, until they find the correct login information.

   The name "brute force" comes from attackers using excessively forceful attempts to gain access to user accounts. Despite being an old cyberattack method, brute force attacks are tried and tested and remain a popular tactic with hackers.

Types of Brute Force Attacks

There are various types of brute force attack methods that allow attackers to gain unauthorized access and steal user data.

a) *Simple Brute Force Attacks:* A simple brute force attack occurs when a hacker attempts to guess a user's login credentials manually without using any software. This is typically through standard password combinations or personal identification number (PIN) codes. These attacks are simple because many people still use weak passwords, such as "password123" or "1234," or practice poor password etiquette, such as using the same password for multiple websites. Passwords can also be guessed by hackers that do minimal reconnaissance work to crack an individual's potential password, such as the name of their favourite sports team.
b) *Dictionary Attacks:* A dictionary attack is a basic form of brute force hacking in which the attacker selects a target, then tests possible passwords against that individual's username. The attack method itself is not technically considered a brute force attack, but it can play an important role in a bad actor's password-cracking process. The name "dictionary attack" comes from hackers running through dictionaries and amending words with special characters and numbers. This type of attack is typically time-consuming and has a low chance of success compared to newer, more effective attack methods.
c) *Hybrid Brute Force Attacks:* A hybrid brute force attack is when a hacker combines a dictionary attack method with a simple brute force attack. It begins with the hacker knowing a username, then carrying out a dictionary attack and simple brute force methods to discover an account login combination. The attacker starts with a list of potential words, then experiments with character, letter, and number combinations to find the correct password. This approach allows hackers to discover passwords that combine common or popular words with numbers, years, or random characters, such as "SanDiego123" or "Rover2020."
d) *Reverse Brute Force Attacks:* A reverse brute force attack sees an attacker begin the process with a known password, which is typically discovered through a network breach. They use that password to search for a matching login credential using lists of millions of usernames. Attackers may also use a commonly used weak password, such as "Password123," to search through a database of usernames for a match.

e) *Credential Stuffing:* Credential stuffing preys on users' weak password etiquette. Attackers collect username and password combinations they have stolen, which they then test on other websites to see if they can gain access to additional user accounts. This approach is successful if people use the same username and password combination or reuse passwords for various accounts and social media profiles.

f) *Cryptanalysis:* Cryptanalysis is the study of methods for obtaining the meaning of encrypted information, without access to the secret information that is typically required to do so. Typically, this involves knowing how the system works and finding a secret key. Cryptanalysis is also referred to as codebreaking or cracking the code. The ciphertext is generally the easiest part of a cryptosystem to obtain and, therefore, is an important part of cryptanalysis. Depending on what information is available and what type of cipher is being analyzed, cryptanalysts can follow one or more attack models to crack a cipher. In this Proposed Algorithm, the encrypted communication data is sent from the sender and that reaches the receiver via different routers, communication channels, and other communication devices. In between this various attackers, crypt analyzers, intruders, and Brute force attacks want to access this secured data by different means and channels. In this process, the intruder alters some of the bits or number of bits in the data packet, which is moving towards the receiver. The proposed work is used learning algorithm K-NN for classification purposes the classification of the sender and then after if the sender is genuine and authenticated. The proposed algorithm's next step is to decrypt the message and after the decryption, a plain text or original message is received by the receiver. This algorithm scans the message and searches the pattern, keywords, sender data, and receiver data, using the digital signature and content of the message previously communicated between the sender and receiver. This algorithm is classifying these data and matches them with their stores' pattern if there is no mismatch is found then the decrypted data or message not altered or fabricated in between the communication path.

If this algorithm finds a mismatch pattern in any of the data patterns which is used in their learning process or pattern, then this algorithm gives the output of a warning message that the security of data is may be breached. The message is altered or distorted by the attacker, crypt analyzer, intruder, and Brute force attacker in between the communication path or channel.

After receiving the warning message from the system, the receiver will communicate with the sender and change their security essential which is used for authentication purposes as well as used for encryption and decryption purposes.

IV. APPLIED METHODOLOGY

The K-nearest neighbour's (KNN) algorithm is a type of supervised ML algorithm that can be used for both classifications as well as regression predictive problems. However, it is mainly used for the classification of predictive problems in the industry.

Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

Working of KNN Algorithm:

K-nearest neighbour's (KNN) algorithm uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of the following steps −

1) Step1 − For implementing any algorithm, we need a dataset. So during the first step of KNN, we must load the training as well as test data.

2) Step2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

3) Step3 − For each point in the test data do the following −

a) Calculate the distance between test data and each row of training data with the help of any of the methods namely: Euclidean, Manhattan, or Hamming distance. The most commonly used method to calculate distance is Euclidean.

b) Now, based on the distance value, sort them in ascending order.

c) Next, it will choose the top K rows from the sorted array.

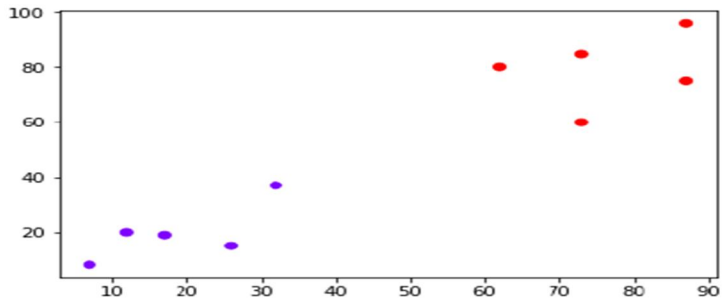
d) Now, it will assign a class to the test point based on the most frequent class of these rows.

4) Step 4 – End

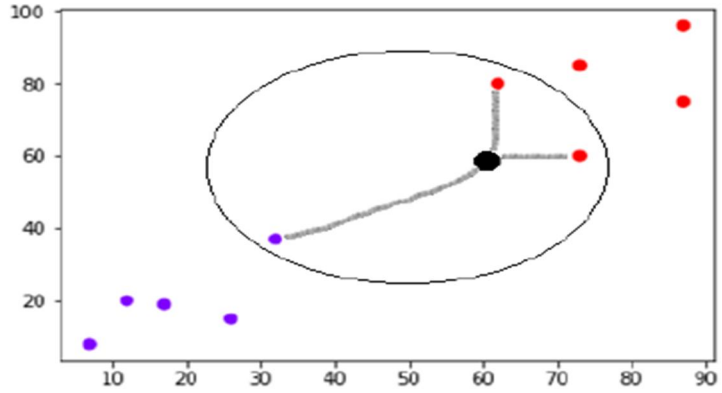
Example

The following is an example to understand the concept of K and the working of the KNN algorithm −

Suppose we have a dataset that can be plotted as follows −



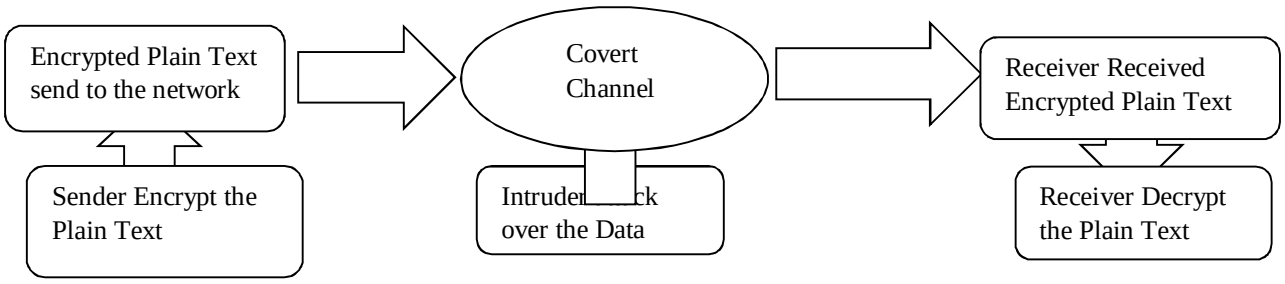
Now, we need to classify new data points with black dots (at points 60,60) into blue or red classes. We are assuming K = 3 i.e. it would find the three nearest data points. It is shown in the next diagram −



We can see in the above diagram the three nearest neighbors of the data point with black dots. Among those three, two of them lie in the Red class hence the black dot will also be assigned to the red class.

Implementation in Python:

As we know K-nearest neighbour's (KNN) algorithm can be used for both classifications as well as regression. The following are the recipes in Python to use KNN as a classifier as well as a regressor.



Working Model of Proposed Algorithm

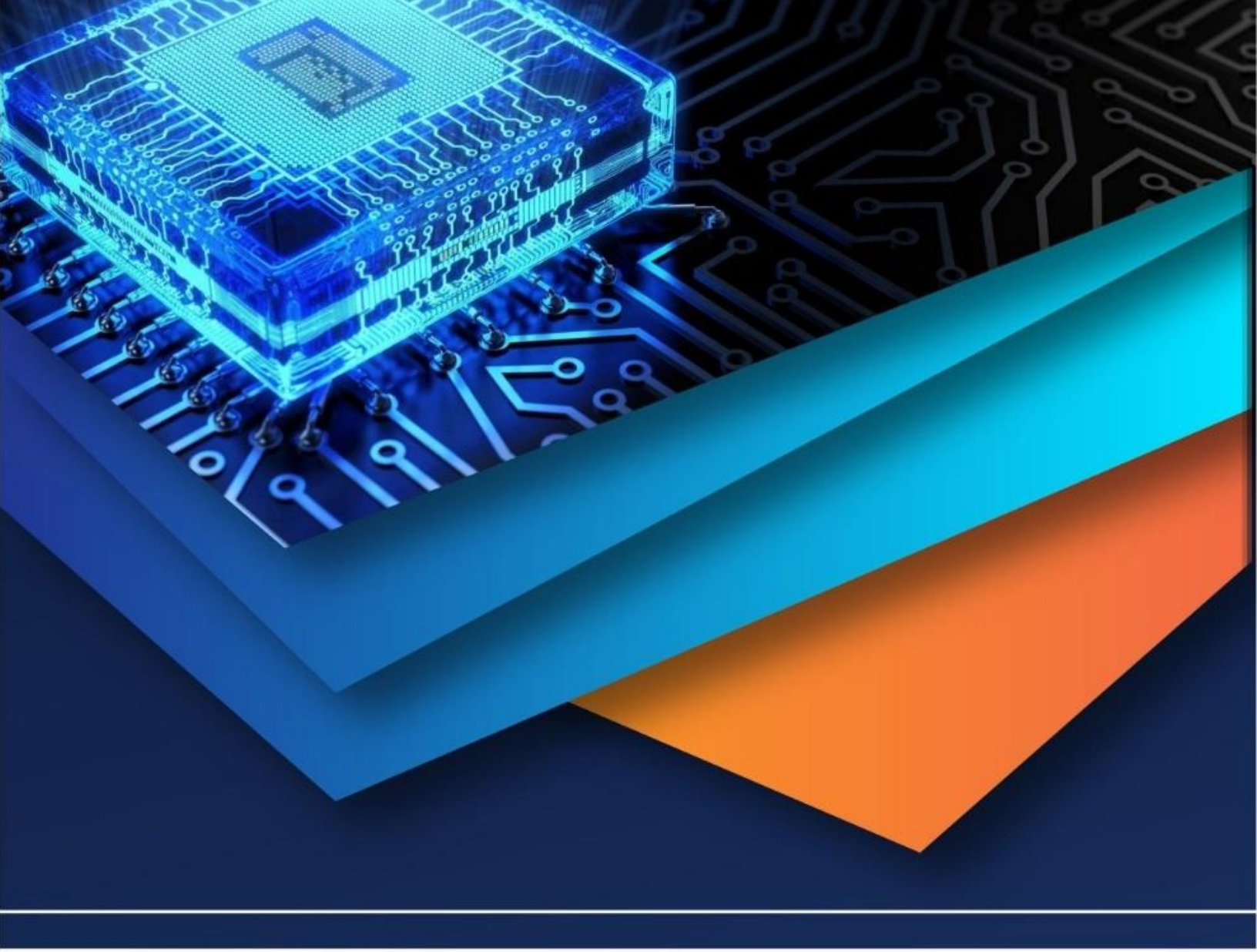
V. CONCLUSION

The fast growth of the internet and the communication world is good means of communication and it has also meant for retrieving information passes in this fastest growing communication network. The data is travel in this channel needs security from the attacker, intruders, crypt analyzer, and brute force attackers. This algorithm enhances the security of data transferred from the sender to receiver by using the K-Nearest Neighbourhood algorithm, this algorithm is used for the classification of data that is coming from previously communicated sender and receiver. Learning from previously communicated message pattern and the content of the message received from the sender to a particular receiver this algorithm learns the pattern and classify the received message is exactly sent by the authenticated sender as well as the content of the message is also not altered or distorted during their path of the communication channel.

REFERENCES

[1] Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab. They describe a focused literature survey of Artificial Intelligence Revised (AI) and Machine learning methods for email spam detection.
[2] K. Agarwal and T. Kumar Harisinghaney et al. (2014) and Mohamad & Selamat (2015) have used the "image and textual dataset for e-mail spam detection with the utilization of assorted methods".
[3] Harisinghaney et al. (2014) have used methods of KNN algorithm and Reversed DBSCAN algorithm with experiments on the dataset. For text recognition, OCR library is employed but this OCR doesn't perform well.
[4] Mohamad & Selamat (2015) uses the feature selection hybrid approach of TF-IDF (Team Frequency Inverse Document Frequency) and Rough pure math.
[5] K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint

Raju Singh Working as an Assistant Professor in the Department of Computer Science and Engineering at Modern Institute of Technology and Research Centre, Alwar, Rajasthan, India. He did B.Tech(CSE) and M.Tech(Information Security) and has 14 years of Teaching Experience in Various reputed Engineering Colleges of the U.P. He had attend International conferences, workshops, Seminars, and published Research Papers in reputed International journals.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY