INTRUSION DETECTION SYSTEM WITH DIMENSIONALITY REDUCTION USING UMAP

Y.Vineela sravya, Kantam Renuka, Kopperala Venkata Swapna, Mahanty Geeta, Madugula Ravali.

Department of Computer Science and Engineering Vignan's Institute of Engineering for women, Duvvada.

ABSTRACT

IDSs and IPSs are critical components in ensuring the security,dependability, and availability of computer networks in businesses. Several intrusion detection systems(IDSs) and intrusion prevention systems (IPSs) have been developed and deployed over the years to improve computer network security.This research focuses on intrusion detection systems(IDSs) that use machine learning(ML) techniques and have demonstrated proficiency and reliability in detecting network attacks. However, the performance of these ML-based IDSs degrades in high-dimensional data spaces.

It is vital to use an appropriate feature extraction strategy that makes use of dimensionality reduction to just get rid of irrelevant features that don't significantly benefit in categorization. Additionally, when trained on unbalanced datasets, ML-based IDSs frequently display poor detection accuracy and a high false positive rate, making it crucial to address this problem during model development. To train and evaluate our algorithms, we analyze the UNSW-NB15 intrusion detection dataset in this paper. Our approach requires using the XG Boost algorithm to implement a filter-based feature reduction approach, which is one of the dimensionality reduction methods,and after wards applying ML techniques like Support Vector Machine(SVM), k-Nearest Neighbor(KNN),Logistic Regression(LR),and Decision Tree(DT) to the condensed features pace for both binary and multiclass classification configurations. The results indicate that the XG Boost-based feature selection technique could enhance the accuracy of our models.Additionally, we suggest using an algorithm like DT to improve the binary classification scheme's accuracy rate.

Keywords: Intrusion Detection, Feature Reduction, Dimensionality Reduction

INTRODUCTION

Hackers' capabilities have evolved at a faster rate as a consequence of the rapid advancement of technology such as the Internet, the Internet of Things (IoT), and communication systems. These individuals are continuously seeking novel methods to compromise computer network security.Intrusion detection systems (IDSs) are now essential parts of computer networks because of how the attack environment is changing. A software or hardware device known as an IDS watches a network for possible dangers and has the ability to respond properly to malicious activity. Host-based IDS (HIDS), distributed IDS (DIDS or NIDS), and hybrid IDS are the three groups into which IDSs fall (HYIDS). Additionally, IDSs are divided into categories based on the methods they use to identify threats, including hybrid, anomaly, and signature-based.Based on their method of discovery, intrusion detection systems (IDSs) can be classified as signature-based, anomaly-based, or hybrid-based. While anomaly-based IDSs watch network activity for suspicious activity and report it, signature-based IDSs scan a database of previously stopped efforts to discover intrusions.IDSs that use a hybrid method incorporate both signature-based and anomaly-based techniques. An IDS's primary design goals are to decrease inaccurate positive alerts and increase detection precision. Machine Learning (ML) based IDSs have become among the most popular of the industry's intrusion detection systems in recent years. Without particular engineering, ML-based apps have the capacity to learn from the past and advance over time.The two primary approaches to machine learning are supervised and unsupervised, with supervised ML utilizing labeled data to build models for binary and multiclass classification tasks. When developing supervised ML models, large datasets with high dimensional feature spaces are frequently used, requiring dimensionality reduction approaches to reduce the number of features needed for training and testing. In this paper, a filter-based feature reduction method utilizing the feature selection measures of the XGBoost Algorithm is considered.It is important to keep in mind that the ML algorithm selected for IDS depends on a number of variables, including the sort of data, the size of the dataset, and the accuracy level required. The XGBoost algorithm, which is a well-known method in the field of machine learning due to its capacity to manage sizable datasets and high-dimensional feature spaces, is utilized in this study for featureselection. A number of measures, including accuracy, precision, and recall, are used to evaluate each ML algorithm's performance.The study's results demonstrated that the ML-based IDSs perform significantly better when using the dimensionality reduction based feature selection method using the XGBoost algorithm, and the decision tree algorithm is found to be the best performing algorithm in the binary classification scheme.

EXISTINGSYSTEM

Even though rule-based systems and other traditional methods have been used in intrusion detection, machine learning techniques have recently shown themselves to perform better. It has been shown that ML algorithms like SVM, kNN, and decision trees can identify intrusions with high accuracy and lesser false positives. Additionally, IDS has used neural networks, including ANN, and has seen outstanding results in detecting the changes. ML methods have shown to produce better outcomes in terms of detection accuracy and reducing the need for human involvement, even though existing techniques have their advantages. It is crucial to notice that some ML algorithms consume more Memory than others, but this can be solved by carefully choosing the right algorithms and right approaches.

Disadvantages

- Inaccuracy
- Requires moretime
- Difficult to manage
- Requires more memory

PROPOSEDSYSTEM

Numerous studies have examined how to evaluate intrusion detection systems, but conventional approaches are often memory and processor time-intensive. In order to solve this problem, our article suggests a feature selection strategy that makes use of the dimensionality reduction provided by the XG Boost algorithm, enhancing accuracy. We create our models using XG Boost in addition to other machine learning methods like Decision Tree (DT),K-Nearest Neighbors(KNN), Logistic Regression (LR), and Support Vector Classifier(SVC). We make use of UMAP to help in visualization.

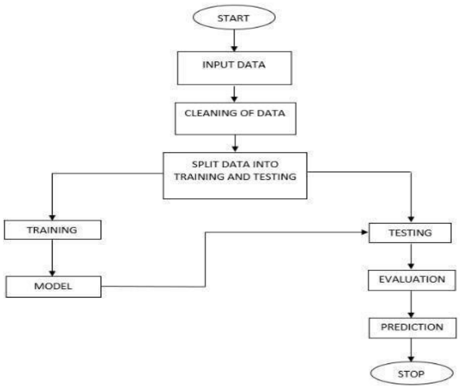


Fig 1:Block Diagram

Advantages

- Time-efficient
- More accurate
- Performs smoothly

ALGORITHMS

XGBoost

Extreme Gradient Boosting, also known as XGBoost, is a distributed method specifically designed to boost speed and efficiency. It focuses on being very effective, adaptable, and portable. It provides a parallel tree boosting technique and implements machine learning algorithms using the Gradient Boosting framework to address a range of data science challenges quickly and accurately. Because of its scalability, XG Boost has recently gained ground and is frequently used to triumph in Kaggle competitions involving applied machine learning and organized data.

Decision Tree

Decision tree learning is a supervised method for building a decision tree from training data in the area of machine learning. A decision tree is a type of predictive model that makes inferences about a target value based on observations about an object. This approach is also known as a reduction tree or a classification tree. In order to build a training model that can predict the group or value of the target variable for new data, decision trees work by learning straightforward decision rules from historical data. In decision trees, we begin at the tree's root to forecast a record's class label. Logistic Regression

A statistical method known as logistic regression is used to forecast binary outcomes, such as "yes" or "no," based on prior data set observations. By examining the relationship between one or more independent variables, a model that predicts a dependent variable is created using this technique. For instance, logistic regression could be used to predict whether a candidate for office would win or lose, or whether a high school student would be accepted into a particular college or rejected. These outcomes' binary nature makes choosing between two options a simple process. As an adaptable tool, logistic regression is frequently used in a variety of disciplines, including economics, medicine, and social sciences.

KNN

One of the simplest machine learning methods based on supervised learning is K-Nearest Neighbor. The K-NN algorithm places the new instance in the category that compares the most favorably to the existing categories, assuming of course that a new case and the existing cases are comparable. This algorithm records all of the information that is accessible and classifies fresh input according to similarity. New data can be accurately and efficiently sorted into the appropriate category using the K-NN method. Numerous applications, including image recognition, recommender systems, and anomaly detection, benefit from this methodology. The K-NN algorithm is a flexible and well-liked machine learning method because it can be modified to handle various data types and distance

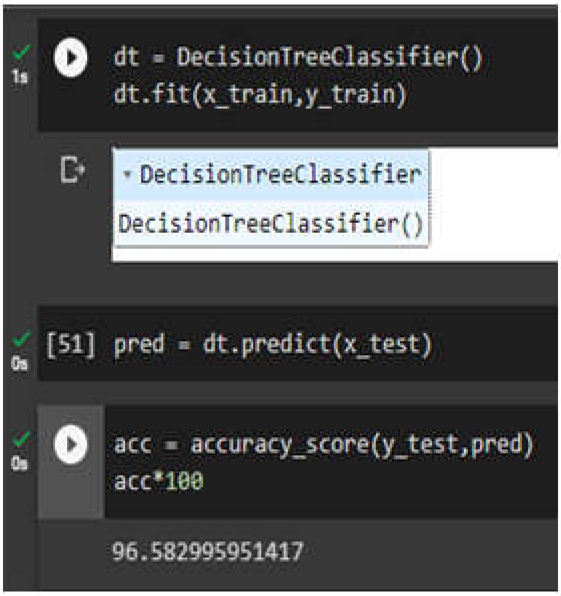
SVC

A popular supervised machine learning method for classification tasks is the Support Vector Classifier (SVC). SVC divides the data into two categories by first mapping the data points to a high-dimensional space and then locating the best hyper plane. The decision boundary between the two categories is known as the hyperplane, and the category with the largest margin is considered to be the best option. To put it another way, the SVC algorithm seeks out the best hyperplane that maximizes the distance between the nearest data points for each class. The data points that are closest to the hyperplane are called support vectors, and they are very important to the algorithm. SVC can accurately classify fresh data points based on the best hyperplane.

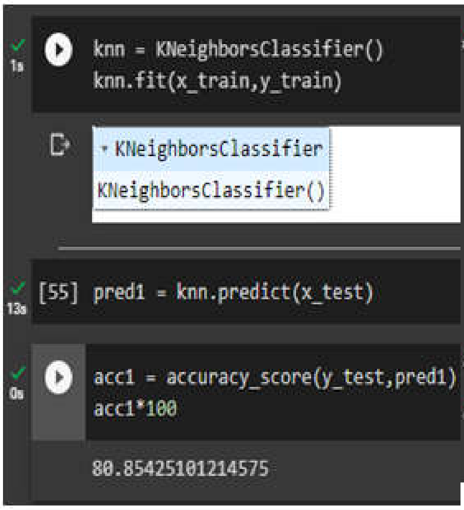
UMAP

The non-linear dimensionality reduction method Uniform Manifold Approximation and Projection (UMAP), or UMAP, employs graph-based techniques to approximate the high-dimensional manifold on which the data is located. It uses this presumption to project the data to a lower-dimensional space while maintaining the local structure of the original data by assuming that the data samples are evenly dispersed across this manifold. By doing this, UMAP is able to preserve more of the data's overall structure than linear dimensionality reduction methods like PCA.

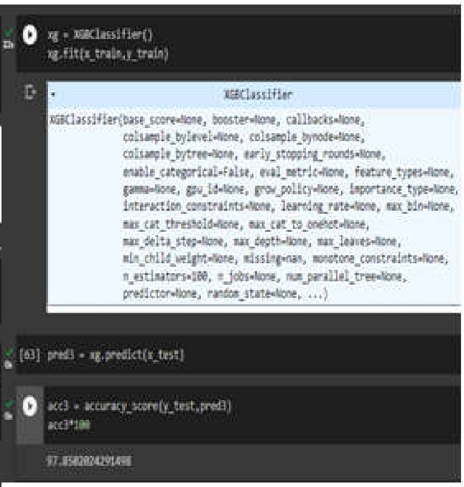
RESULTS



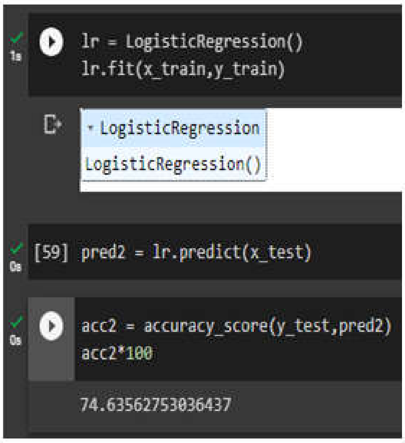
```
dt = DecisionTreeClassifier()
dt.fit(x_train,y_train)

▾ DecisionTreeClassifier
DecisionTreeClassifier()

[51] pred = dt.predict(x_test)

acc = accuracy_score(y_test,pred)
acc*100

96.582995951417
```



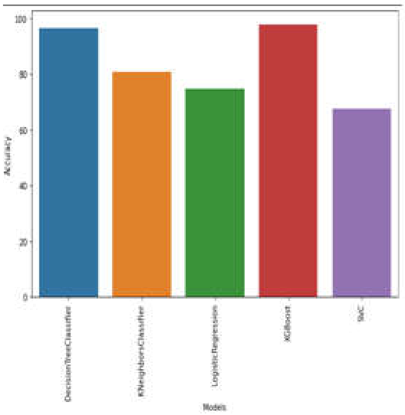
```
knn = KNeighborsClassifier()
knn.fit(x_train,y_train)

▾ KNeighborsClassifier
KNeighborsClassifier()

[55] pred1 = knn.predict(x_test)

acc1 = accuracy_score(y_test,pred1)
acc1*100

80.85425101214575
```

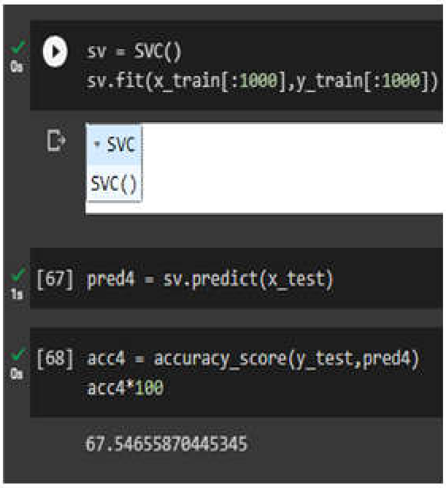


```
xg = XGBClassifier()
xg.fit(x_train,y_train)

▾ XGBClassifier
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, ...)

[63] pred3 = xg.predict(x_test)

acc3 = accuracy_score(y_test,pred3)
acc3*100

97.8582024291498
```

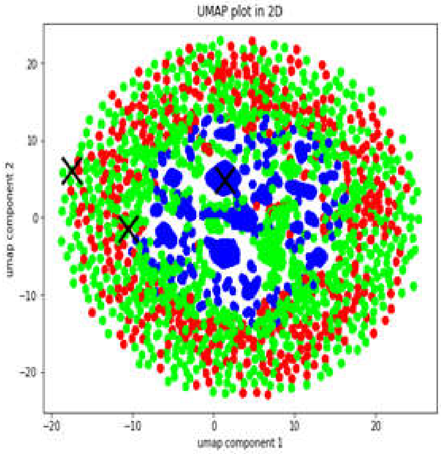


```
lr = LogisticRegression()
lr.fit(x_train,y_train)

▸ LogisticRegression
LogisticRegression()

[59] pred2 = lr.predict(x_test)

acc2 = accuracy_score(y_test,pred2)
acc2*100

74.63562753036437
```





```
sv = SVC()
sv.fit(x_train[:1000],y_train[:1000])

▸ SVC
SVC()

[67] pred4 = sv.predict(x_test)

[68] acc4 = accuracy_score(y_test,pred4)
acc4*100

67.54655870445345
```



CONCLUSION

In order to handle the database and improve the accuracy of the intrusion detection system applying the XGBoost algorithm, this study uses dimensionality reduction techniques. The models are also trained using a variety of machine learning techniques, including decision trees, K-Nearest Neighbor, logistic regression, and SVC. The efficiency of the machine learning algorithms is measured using performance indicators including , accuracy, precision and recall. Through the application of dimensionality reduction techniques, our research provides improved accuracy results when compared to earlier studies.

REFERENCE

1. Ribeiro J, Saghezchi FB, Mantas G,Rodriguez J, Abd-Alhameed RA.Hidroid: prototyping a behavioral host-based intrusion detection and prevention system for android. IEEEAccess. 2020 ; 8 :23154–168.
2. Van NTT, ThinTN. Accelerating anomaly-based IDS using neural networkonGPU. In:2015 international conference one advanced computing and applications(ACOMP). IEEE; 2015. pp. 67–74.
3. jabez J, Muthukumar B. Intrusion detection system (IDS): anomaly detection using outlier detection approach.ProcediaComputSci.2015;48:338–46.
4. Neelakantan S,Rao S.A threat aware anomaly based intrusion detection approachfor obtaining network specific useful alarms.
In: International conference on distributed computing anetworking.Springer.2009;