

Enhancing Intrusion Detection Systems with Explainable AI for False Positive Identification

Dr. Venkata Reddy Medikonda, Associate Professor in the Department of Computer Science and Engineering,
Marri Laxman Reddy Institute of Technology and Management, Hyderabad, Telangana, India

Dr T.S.Sreenivas, Associate Professor, Department of CSE , MLRITM, tumulurisri@mlritm.ac.in

Dr. N.Pushpalatha, Associate professor in CSE, Marri Laxman Reddy Institute of Technology and Management,
Hyderabad, Dindigul

ABSTRACT: This work means to utilize “eXplainable Artificial Intelligence (XAI)” strategies to address false positives in intrusion detection systems. Working on the interpretability of AI calculations assists with explaining and uncover the dynamic interaction. The technique depends on the idea that the exactness of interruption discovery relates with the significance, not entirely set in stone by XAI strategies. The examination is to test the reasonable pertinence and generalizability of the proposed strategy to genuine online protection circumstances through the LYCOS-IDS2017 dataset for trial and error. The outcomes feature the potential benefits of including XAI techniques into frameworks of interruption location. This shows, by doing this, XAI might give a more exact and interpretable technique for spotting and diminishing bogus up-sides in network protection. This coordination presents a possible way to work on the overall proficiency of interruption identification frameworks in down to earth utilizes.

INDEX TERMS Intrusion detection, machine learning, explainability, XAI, false positive rate

1. INTRODUCTION:

The security of PC frameworks is main goal in the connected advanced landscape of today to safeguard private data and assurance constant activities. Among the numerous security frameworks utilized, interruption location is an essential cautious system as it is the last line of insurance against security strategy infringement and endeavors of unlawful access. As Stallings and Brown [1] call attention to, interruption discovery is fundamental in saving framework respectability and security and is in this way fairly significant in the bigger system of online protection.

Intended to find and respond to unsafe way of behaving or strategy infractions inside a PC organization or framework, "intrusion detection systems (IDS)" are These frameworks distinguish potential risks through example and conduct examination of organization traffic or framework movement. Two fundamental classes could assist one with for the most part understanding the discovery cycle: peculiarity based identification and mark based recognition [2].

Signature-based discovery finds unfriendly exercises utilizing laid out examples or marks of known attacks. Normally created from earlier attacks or known weaknesses, these marks let the

IDS distinguish and respond to specific examples connected with pernicious action. In spite of the fact that signature-based identification is really great for spotting known dangers, its reliance on foreordained marks confines its ability to see as new or before neglected attacks.

On the other hand, oddity based identification searches for takeoffs from anticipated conduct inside a framework or organization. Dissimilar to signature-based location, irregularity put together techniques don't depend with respect to pre-characterized assault designs. Rather, they give a benchmark of commonplace way of behaving and show varieties from this benchmark as potential perils. Alazab et al. [3] bring up that this approach has an advantage in spotting until now unidentified or zero-day attacks. Be that as it may, regularly contrasted with signature-based frameworks, peculiarity based identification delivers all the more misleading up-sides [4].

In interruption discovery frameworks, bogus up-sides — that is, misleading distinguishing pieces of proof of harmless action as dangers — cause an extraordinary trouble. Misleading alarms in the IDS could cause client disturbance, framework overburden, and trust misfortune [5]. Successful working of interruption location frameworks relies upon lessening bogus up-sides so.

Misleading up-sides in interruption location frameworks have been proposed to be tended to utilizing many methodologies. One procedure is working on the exactness of location calculations by utilization of modern AI approaches or component determination methodologies [6]. Moreover assisting with bringing down misleading problems and increment framework general steadfastness are post-handling techniques incorporate alarm

separating, copy discovery, and gathering like identifications [7].

Utilizing Logical Man-made consciousness (XAI) advances to work on the interpretability and constancy of interruption identification frameworks has drawn in expanding consideration as of late. XAI chips away at making simulated intelligence models and calculations that can give human clients contentions or clarifications for their decisions, subsequently explaining the dynamic cycle [8]. Scientists need to increment general exactness and steadfastness by coordinating XAI strategies into interruption discovery frameworks, accordingly empowering the framework to recognize genuine perils and bogus up-sides.

In this venture, the recommended technique is utilizing XAI strategies to settle misleading up-sides in interruption discovery frameworks. The framework can proficiently recognize and decrease bogus location by learning the significance of highlights utilizing XAI strategies and coordinating this data with the certainty estimations of the calculation. Through genuine world datasets, the task looks to affirm the down to earth convenience and viability of this strategy, hence supporting the ceaseless endeavors to work on the security and reliability of interruption recognition frameworks.

We will examine the troubles introduced by misleading up-sides in interruption discovery frameworks, explore the potential benefits of including XAI techniques into these frameworks, and depict the methodology and objectives of the proposed project in the accompanying areas.

LITERATURE SURVEY

Inside the field of network safety, interruption location is exceptionally imperative for shielding organizations and PC frameworks from unlawful

access endeavors and antagonistic way of behaving. From signature-based location to more modern AI techniques, specialists have researched multiple ways throughout the years to expand the reliability and adequacy of intrusion detection systems (IDS). With an eye on the challenges introduced by misleading up-sides and the potential benefits of eXplainable Artificial Intelligence (XAI) instruments, this writing survey offers a rundown of current work on interruption identification.

The improvement of interruption discovery frameworks is researched in a careful survey by Nisioti et al. [1] alongside the shift from ordinary procedures to additional complex ones consolidating solo learning draws near. The authors pressure the need of assailant attribution in interruption location and go over the capability of unaided strategies in spotting uncommon action liberated from dependence on foreordained marks. Understanding the territory of interruption discovery and the troubles with attribution and oddity identification relies much upon this review.

With an eye on the restrictions of regular mark based strategies, Sommer and Paxson [2] investigate the utilization of AI methods for network interruption location. The shut world suspicion incorporated into signature-based frameworks, the creators fight, couldn't be adequate to mirror the intricacy and assortment of genuine attacks. Stressing the chance of these techniques to work on the adequacy of interruption recognition frameworks, they support the work of AI calculations to recognize up until recently imperceptible or zero-day attacks.

Utilizing an ill-disposed technique for Logical artificial intelligence in interruption location frameworks, Marino et al. [3] propose in a paper The ill-disposed structure makes it conceivable to make unfriendly cases that might be explored to uncover

blemishes and look at the IDS's flexibility. Specialists might find out about the IDS's dynamic cycle and point out any blemishes by fluctuating info information and global positioning framework response. This technique assists with making major areas of strength for progressively open interruption identification frameworks.

Looking at the issue of "for what reason would it be advisable for me I trust you?" Ribeiro et al. [4] give an approach to explaining the expectations of AI classifiers. By approximating the way of behaving of the classifier in the space of the occurrence of interest, the creators offer a technique known as LIME (Nearby Interpretable Model-skeptical Clarifications), which produces clarifications for individual expectations. This technique assists clients with assessing the classifier's trustworthiness and handle the components causing a specific expectation. LIME works on the receptiveness and believability of AI models — incorporating those utilized in interruption recognition — by offering reasonable clarifications.

Utilizing spread of actuation contrasts, Shrikumar et al. [5] give a strategy to learning critical highlights in brain organizations. Through network initiations both with and without the component present, the creators recommend a method known as "DeepLIFT (deep Learning Significant Elements)", which processes the commitment of each and every info component to the result expectation. This strategy assists with recognizing significant components affecting the dynamic course of the model, accordingly empowering interpretability and understanding of complex brain network models.

The writing survey stresses commonly the need of incorporate Reasonable man-made intelligence techniques and the need of adjusting bogus up-sides in interruption discovery frameworks. Utilizing XAI

procedures and AI calculations assists scientists with expanding the interpretability, straightforwardness, and constancy of interruption location frameworks, subsequently further developing the security stance of PC organizations and frameworks.

2. **METHODOLOGY**

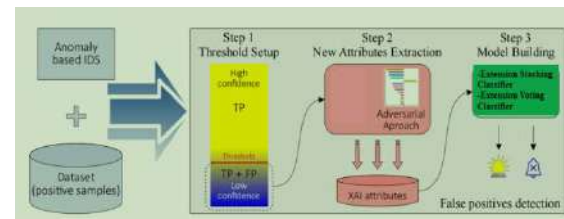
a) **Proposed work:**

eXplainable Artificial Intelligence (XAI) techniques, the proposed framework tries to diminish bogus up-sides in interruption location. The framework works on its ability to track down bogus up-sides by accumulating the certainty estimations of the calculation with property pertinence obtained through XAI procedures. This technique ensures the exactness of genuine positive discoveries while ensuring harmless exercises are not wrongly recognized as dangers. Through this XAI reconciliation, the framework gives more straightforwardness and interpretability, consequently empowering clients to get a handle on the dynamic interaction and depend on the results [1]. This proposed approach upholds the consistent drives targeting working on the constancy and productivity of interruption location frameworks in useful network protection conditions.

b) **System Architecture:**

The recommended peculiarity based IDS framework configuration begins with network information consumption from a dataset, in this manner offering a total image of framework movement. Preprocessing of this material ensures consistency and utilization. A limit setting stage then sets healthy levels of takeoff from common way of behaving, thus coordinating peculiarity ID. New characteristics extraction further develops the element space, in this manner expanding the dataset with significant data. Preparing AI calculations to

find irregularities relying upon procured highlights and set edges is the substance of model turn of events. Bogus up-sides location limits phony problems by deciphering model decisions utilizing eXplainable Artificial Intelligence (XAI) techniques, in this way ensuring solid distinguishing proof of genuine dangers. Fundamental for productive network safety strategies, this plan advances an unmistakable and justifiable interruption identification system [1].



“Fig 1 Proposed Architecture”

c) **Dataset collection:**

Assessing intrusion detection systems (IDS) and researching network safety issues benefit much from the LYCOS-IDS2017 dataset. This dataset incorporates a wide range of organization traffic information accumulated from a few sources, including both customary movement and reenacted attacks. Accumulated over an imperative timeframe, the dataset offers a total image of real organization action, permitting scholastics to look at and handle the subtleties of digital risks. The LYCOS-IDS2017 dataset upholds administered learning strategies for IDS appraisal and benchmarking utilizing marked occasions meaning harmless and antagonistic action. Its intensive inclusion of many assault sorts and organization conventions likewise makes it conceivable to make and test solid interruption location frameworks. Approaching this data empowers scholastics to research imaginative thoughts for further developing organization

security and along these lines lessening digital risks [1].

A	B	C	D	E	F	G	H	I
Dst Port	Protocol	Timestamp	Flow Durati	Tot Fwd Pk	Tot Bwd Pk	TotLen	FwdTotLen	BwdTotLen
0	0	14-02-2018 08:31	112641719	3	0	0	0	0
0	0	14-02-2018 08:32	112641466	3	0	0	0	0
0	0	14-02-2018 08:36	112639623	3	0	0	0	0
22	6	14-02-2018 08:40	6083906	15	10	1239	2273	744
22	6	14-02-2018 08:40	6084006	14	11	1143	2209	744
22	6	14-02-2018 08:40	6085341	16	12	1239	2273	744
0	0	14-02-2018 08:29	112640480	3	0	0	0	0
0	0	14-02-2018 08:52	112641284	3	0	0	0	0
RD	6	14-02-2018 08:47	478513	5	3	211	463	211
RD	6	14-02-2018 08:47	475648	5	3	220	472	209
RD	6	14-02-2018 08:47	474926	5	3	220	472	209
RD	6	14-02-2018 08:47	472471	5	3	209	461	209
RD	6	14-02-2018 08:47	512758	5	3	211	463	211
RD	6	14-02-2018 08:47	476711	5	3	206	458	206
RD	6	14-02-2018 08:47	476616	5	3	211	463	211
RD	6	14-02-2018 08:47	477101	5	3	211	463	211
RD	6	14-02-2018 08:47	474676	5	3	214	466	214
RD	6	14-02-2018 08:47	476608	5	3	200	461	200
RD	6	14-02-2018 08:47	476089	5	3	215	467	215
RD	6	14-02-2018 08:47	473957	5	3	215	467	215

“Fig 2 data set”

d) **DATA PROCESSING**

Involving the pandas module in Python, information handling for the LYCOS-IDS2017 dataset handles the information really inside a dataframe structure. The dataset is initial placed into a pandas dataframe so information examination and change might be done without any problem. To streamline the dataset and increment processing execution, undesirable segments — like repetitive or trivial elements — are erased from the dataframe. This stage ensures just significant elements vital for interruption location investigation are kept. Information pretreatment tasks like cleaning, changing over, and putting together the dataset might be really finished by utilizing the adaptability and utility of pandas. Eventually, this step of information handling prepares the dataset for include designing, model development, and further examination, hence empowering the production of effective interruption identification frameworks ready to definitively distinguish and decrease digital weaknesses [1].

e) **VISUALIZATION**

Seaborn and Matplotlib envisioning serious areas of strength for offers for investigating the LYCOS-IDS2017 dataset. While Matplotlib gives a flexible climate to creating custom tailored representations, Seaborn offers an undeniable level point of

interaction for planning engaging measurable visuals. < Utilizing these libraries permits one to explore graphically numerous aspects of the dataset, including network traffic dispersion, assault occurrence examples, and element relationships. Specialists might make instructive diagrams like histograms, disperse plots, heatmaps, and bar graphs utilizing Seaborn's underlying factual perception devices and Matplotlib's far reaching modifying decisions. These portrayals support information on the essential information structure, spot any examples or abnormalities, and guide further exploration and model structure. In network protection, perception using Seaborn and Matplotlib at long last works on the interpretability and coherence of results, consequently supporting information driven decision-production [1].

f) **Feature Selection**

Assembling great interruption recognition frameworks relies fundamentally upon highlight choice. One technique for highlight determination is "Mutual Information Classification (MIC)". Suitable for deciding valuable elements for characterization issues, common data measures the factual dependence between two factors. Holding just an assigned level of the greatest scoring highlights, SelectPercentile picks the best elements in light of their scores from Common Data Grouping. This approach diminishes the dimensionality of the dataset by killing copy or trivial components, hence saving the most significant data for interruption recognition. Since it focuses on the characteristics probably going to help legitimate arrangement, this technique works on the effectiveness and viability of later model creation. By giving the most discriminative highlights first concern,

include choice utilizing SelectPercentile with Shared Data Order expands the interruption recognition framework execution [1].

h) Lable encoding

Many AI procedures need preprocessing in which case mark encoding utilizing LabelEncoder changes class names into mathematical portrayals. Inside the structure of interruption discovery, absolute names like "ordinary" and "assault" should be switched into mathematical qualities over completely to help model preparation. From the scikit-learn bundle, LabelEncoder really finishes this task. It gives each particular name in the dataset a singular number, subsequently changing over all out factors into a configuration AI models can without much of a stretch handle and handle. LabelEncoder ensures consistency and fit with downstream investigation and model structure processes by consistently encoding marks across the dataset. Name encoding assists with setting up the dataset for additional examination and model preparation in the structure of bogus positive distinguishing proof in interruption recognition using eXplainable Artificial Intelligence (XAI), subsequently empowering the development of exact and interpretable interruption identification frameworks [1].

i) TRAINING AND TESTING

Inside the system of misleading positive ID in interruption location utilizing eXplainable Artificial Intelligence (XAI), the preparation and testing strategy comprises of numerous significant stages. Typically utilizing strategies like cross-valuation to ensure power and generalizability of the model, the dataset first is parted into preparing and testing sets. AI calculations are shown on named information during preparing involving XAI strategies to further

develop interpretability and receptiveness in the dynamic cycle. SelectPercentile with Shared Data Arrangement and other component determination procedures might be utilized to find informational components decreasing overfitting. The model's presentation on the testing set subsequent to preparing is surveyed to check whether it can accurately detect misleading up-sides and separate harmless from hurtful action. The forecasts of the model are then analyzed utilizing XAI strategies to acquire comprehension of the components prompting misleading up-sides and help the interruption discovery framework to be gotten to the next level. This iterative methodology ensures the making of predictable and interpretable models for interruption location bogus positive ID [1].

j) ALGORITHMS:

Random FOrEst

Building numerous choice trees during preparing and creating the method of the classes as the expectation, "Random Forest (RF)"[8] is an outfit learning strategy. In our examination, RF is utilized for interruption identification, in which it joins the expectations of numerous choice trees to further develop exactness and versatility in spotting destructive movement while bringing down misleading up-sides.

KNN

A non-parametric characterization strategy, K-nearest Neighbors (KNN[9]) names a piece of information concurring on the greater part class of its k nearest neighbors. In our exploration, KNN is utilized for interruption identification to classify network traffic information by surveying the closeness among events and their neighbors, accordingly assisting with spotting irregularities and potential perils.

Decision Tree

A non-parametric characterization technique, K-nearest Neighbors (KNN[9]) marks a significant piece of information concurring on the greater part class of its k nearest neighbors. In our examination, KNN is utilized for interruption recognition to arrange network traffic information by surveying the similitude among events and their neighbors, hence assisting with spotting anomalies and potential risks.

Naive Bayes

In light of Bayes' hypothesis with the assumption of element freedom, Guileless Bayes[11] is a probabilistic grouping framework. Gullible Bayes is utilized in our undertaking for interruption recognition to gauge the probability of a given case having a place with a specific class, in this manner assisting with distinguishing potential dangers and misleading up-sides relying upon the noticed information.

Neural Network

Including connected layers of neurons, neural Network[12] is an adaptable AI model propelled by the design and capability of the human cerebrum. In our review, complex examples and relationships in network traffic information are picked up utilizing brain organizations, thusly permitting dependable arrangement of ordinary and unsafe action utilizing high-layered highlight spaces.

Voting Classifier - RF + AB

Joining numerous autonomous classifiers, Casting a ballot Classifier [13] makes larger part casting a ballot expectations. Utilizing the qualities of each base classifier to increment general execution and constancy in spotting bogus up-sides and potential dangers, our review utilizes a Democratic Classifier made of Irregular Backwoods and AdaBoost classifiers for interruption identification.

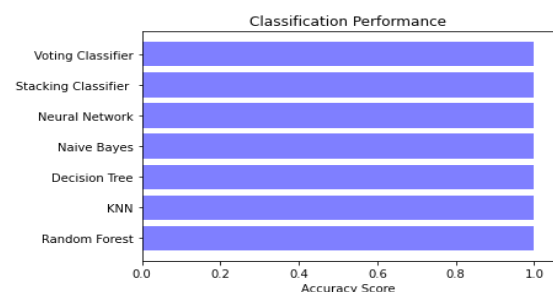
Stacking Classifier - RF + MLP with LightGBM

Joining numerous classifiers involving a meta-classifier helps Stacking Classifier[14] give expectations in troupe learning. Utilizing the different qualities of each base classifier and the meta-classifier to further develop precision and strength in recognizing misleading up-sides and malignant movement, our task utilizes a Stacking Classifier containing Irregular Backwoods, Multi-layer Perceptron (MLP), and LightGBM classifiers for interruption discovery.

3. EXPERIMENTAL RESULTS

Accuracy: The limit of a test to precisely isolate the wiped out from the solid cases characterizes its exactness. Computing the level of genuine positive and genuine negative in undeniably examined occurrences will assist us with assessing the exactness of a test. Numerically, this is communicated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



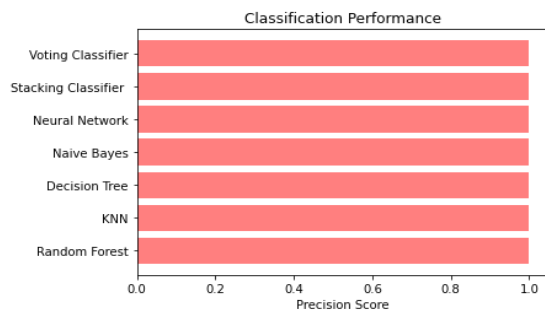
“Fig 3 ACCURACY COMPARISON GRAPH”

Precision: Accuracy measures among the ones arranged as up-sides the extent of appropriately

recognized occasions or tests. Thus, the recipe to get the exactness is:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} = \frac{TP}{TP + FP}$$

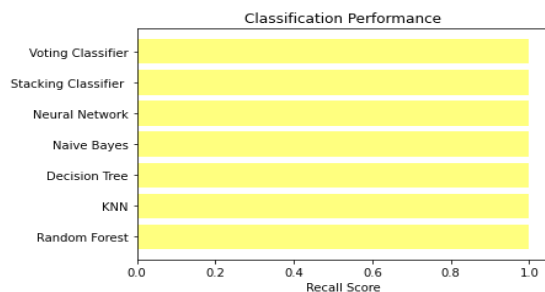
$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$



“Fig 4 PRECISION COMPARISON GRAPH”

Recall: In AI, review is a measurement checking a model's ability to track down all significant occurrences of a given class. It shows the fulfillment of a model as far as precisely anticipated positive perceptions to add up to genuine up-sides, thusly directing comprehension of this perspective.

$$\text{Recall} = \frac{TP}{TP + FN}$$

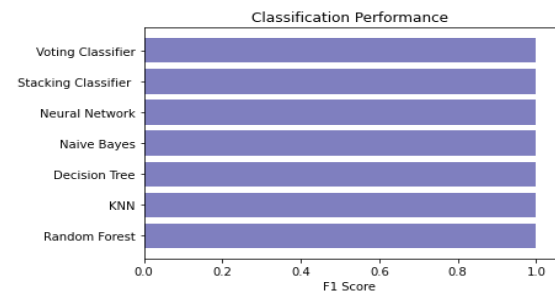


“Fig 5 RECALLCOMPARISON GRAPH”

F1-Score: In AI, F1 score is a measurement of model rightness. It mixes a model's review and exactness scores. Across the entire dataset, the exactness measure counts the times a model delivered a right expectation.

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



“FIG 6 F1 SCORE COMPARISON GRAPH”

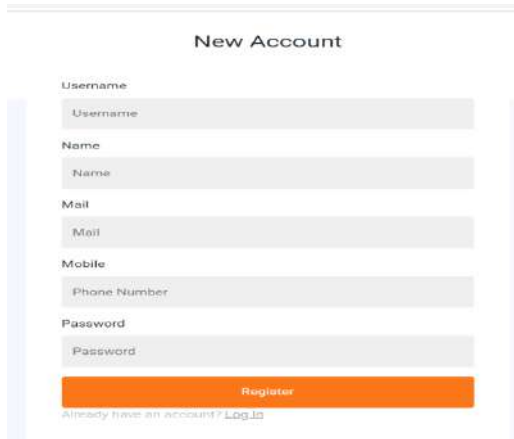
ML Model	Accuracy	f1_score	Recall	Precision
Random Forest	0.871	0.869	0.871	0.897
KNN	0.871	0.869	0.871	0.897
Decision Tree	0.871	0.869	0.871	0.897
Naive Bayes	0.566	0.598	0.566	0.863
Neural Network	0.871	0.869	0.871	0.897
Extension Stacking Classifier	0.871	0.869	0.871	0.897
Extension Voting Classifier	1.000	1.000	1.000	1.000

“Fig 7 PERFORMANCE EVALUATION”

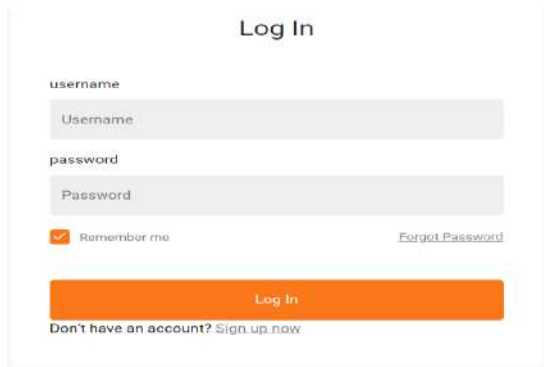
TABLE



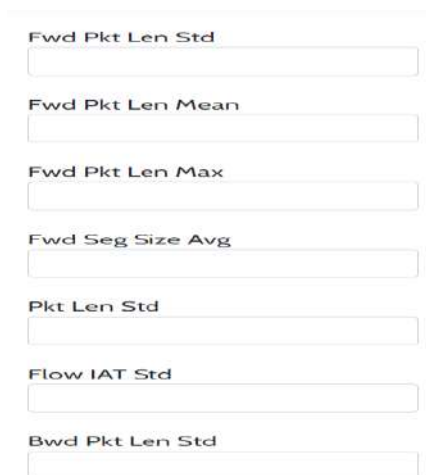
“FIG8 HOME PAGE”




“FIG 9 SIGN UP PAGE”



“FIG 10 SIGNIN PAGE”



“FIG11 UPLOAD INPUT DATA”



“FIG 12 UPLOAD INPUT DATA “

Result: There is an Attack Detected, Attack Type is DDoS!

”FIG 13 PREDICTED RESULT”

Result: There is an No Attack Detected, it is Normal!

”FIG 14 PREDICTED RESULT”

5. CONCLUSION

At last, this investigation underscores how critically "eXplainable Artificial Intelligence (XAI)" approaches ought to be utilized to further develop "intrusion detection system (IDS)" execution. The undertaking accomplishes a more refined dynamic cycle by investigating the main drivers of misleading up-sides and joining quality importance removed from XAI with certainty estimates delivered by the IDS calculation, so further developing exactness in isolating between bogus location and genuine dangers.

The improvement of a strengthening AI calculation assists with bettering handle the connection between XAI-produced highlights and location validity, consequently working with the tweaking of the IDS's choice cutoff points. With an unmistakable

diminishing in bogus up-sides and least misfortune in obvious up-sides, assessment utilizing the LYCOS-IDS2017 dataset shows the pragmatic ease of use and value of the proposed technique.

To reinforce framework security by further developing exactness, trustworthiness, and ease of use, the venture's general objective eventually rises above misleading positive decrease to incorporate an exhaustive redesigning of IDS. Through these drives, the venture significantly assists with propelling the discipline of interruption discovery and reinforce general network safety approaches.

6. FUTURE SCOPE

This exploration is to investigate and incorporate further developed approaches took special care of organization traffic qualities, accordingly propelling the reconciliation of “eXplainable Artificial Intelligence (XAI)” procedures into interruption location models. This drive will work on model interpretability and dependability, thusly giving security experts better comprehension of framework dynamic strategies. To bring down misleading up-sides much further, the task expects to explore and analyze refined AI models for interruption identification and assess calculations ready to grasp complex examples in network information. Continuous utilization of the recommended approach in interruption location frameworks is additionally imagined, requiring beating obstructions connected with processing proficiency for perfect combination into functional security frameworks. Besides, dynamic variation of the interruption identification framework to new cyberthreats is fundamental and requires consistent checking and refreshing of the XAI-upgraded model to give strength against creating perils in powerful organization settings.

REFERENCES

- [1] A. M. Riyad, M. Ahmed, and H. Almistarihi, “A quality framework to improve ids performance through alert post-processing,” *International Journal of Intelligent Engineering and Systems*, 2019.
- [2] R. Alshammari, S. Sonamthiang, M. Teimouri, and D. Riordan, “Using neuro-fuzzy approach to reduce false positive alerts,” in *Fifth Annual Conference on Communication Networks and Services Research (CNSR'07)*, pp. 345–349, 2007.
- [3] A. Nisioti, A. Mylonas, P. D. Yoo, and V. Katos, “From intrusion detection to attacker attribution: A comprehensive survey of unsupervised methods,” *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3369–3388, 2018.
- [4] K. A. Scarfone and P. M. Mell, “Sp 800-94. guide to intrusion detection and prevention systems (idps),” tech. rep., National Institute of Standards & Technology, Gaithersburg, MD, USA, 2007.
- [5] R. Sommer and V. Paxson, “Outside the closed world: On using machine learning for network intrusion detection,” in *2010 IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.
- [6] E. K. Viegas, A. O. Santin, and L. S. Oliveira, “Toward a reliable anomaly-based intrusion detection in real-world environments,” *Computer Networks*, vol. 127, pp. 200–216, 2017.
- [7] Internet Steering Committee project in Brazil, “Total data traffic on the brazilian internet,” 2022. <https://ix.br/agregado/>. Accessed on: Nov. 11, 2022.
- [8] D. L. Marino, C. S. Wickramasinghe, and M. Manic, “An adversarial approach for explainable ai

in intrusion detection systems,” in IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society, pp. 3237–3243, 2018.

[9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[10] L. S. Shapley, *A Value for n-Person Games*, pp. 307–317. Princeton University Press, 1953.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should i trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.

[12] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not just a black box: Learning important features through propagating activation differences,” *ArXiv*, vol. abs/1605.01713, 2016.

[13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLOS ONE*, vol. 10, pp. 1–46, 07 2015.

[14] MIT Lincoln Laboratory, “1999 darpa intrusion detection evaluation dataset,” 1999. <https://www.ll.mit.edu/r-d/datasets/1999-darpa-intrusion-detection-evaluation-dataset>. Accessed on: Nov. 16, 2022.

[15] G. P. Spathoulas and S. K. Katsikas, “Reducing false positives in intrusion detection systems,” *Computers & Security*, vol. 29, no. 1, pp. 35–44, 2010.

[16] P. Pitre, A. Gandhi, V. Konde, R. Adhao, and V. Pachghare, “An intrusion detection system for zero-day attacks to reduce false positive rates,” in *2022 International Conference for Advancement in Technology (ICONAT)*, pp. 1–6, 2022.

[17] H. Kim, Y. Lee, E. Lee, and T. Lee, “Cost-effective valuable data detection based on the reliability of artificial intelligence,” *IEEE Access*, vol. 9, pp. 108959–108974, 2021.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” 2017.

[19] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, “A survey of network-based intrusion detection data sets,” *Computers & Security*, vol. 86, pp. 147–167, 2019.

[20] G. Engelen, V. Rimmer, and W. Joosen, “Troubleshooting an intrusion detection dataset: the cicsids2017 case study,” in *2021 IEEE Security and Privacy Workshops (SPW)*, pp. 7–12, 2021.

[21] A. Rosay, E. Cheval, F. Carlier, and P. Leroux, “Network Intrusion Detection: A Comprehensive Analysis of CIC-IDS2017,” in *8th International Conference on Information Systems Security and Privacy*, pp. 25–36, SCITEPRESS - Science and Technology Publications, Feb. 2022.

[22] M. Ring, A. Dallmann, D. Landes, and A. Hotho, “IP2Vec: Learning similarities between ip addresses,” in *2017 IEEE International Conference*

on Data Mining Workshops (ICDMW), pp. 657–666, 2017.

[23] A. Nisioti, A. Mylonas, P. D. Yoo, and V. Katos, “From intrusion detection to attacker attribution: A comprehensive survey of unsupervised methods,” *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3369–3388, 2018.

[24] R. Sommer and V. Paxson, “Outside the closed world: On using machine learning for network intrusion detection,” in *2010 IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.

[25] D. L. Marino, C. S. Wickramasinghe, and M. Manic, “An adversarial approach for explainable ai in intrusion detection systems,” in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, pp. 3237–3243, 2018.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should i trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.

[27] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not just a black box: Learning important features through propagating activation differences,” *ArXiv*, vol. abs/1605.01713, 2016.