Enhancing a Dataset to Improve Anomaly Detection Using Machine Learning

V. Sankar, Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India
Dr. G. Zayaraz, Professor, Department of CSE, Pondicherry Engineering College, Pondicherry

ABSTRACT This paper discusses the data security of customer and business data. The business website receives requests from various sources such as customers, crawlers, bots, and hackers. The elimination of anomalies improves the effective usage of hardware, software, and network bandwidth. A machine learning technique is used to identify anomalies from each API request.

The quality of the dataset is important for better accuracy. This paper proposes to improve the quality of the training dataset using Quadratic Discriminant Analysis and Linear Discriminant Analysis models. The proposed method is expected to yield better accuracy, precision, and F1-score.

Keywords – Quadratic Discriminant Analysis and Linear Discriminant Analysis, Web service security, Anomaly detection, Machine Learning.

I. INTRODUCTION

E-Commerce security ensures customer data are stored safely and all transactions are carried out without any compromise in data security. Cyber attacks on e-commerce may come in many forms and proper preventive measures are to be taken to ensure data protection. Some common types of attacks are phishing, monetary theft, credit card fraud, hacking, and misusing of intellectual property.

Any e-commerce platform ensures proper protection against these threats and prioritize security measures for its customers. The machine learning mechanism is used to restrict anomaly access.

Analyzing datasets using machine learning can help in processing huge datasets and detect patterns that could be used to isolate attacks and malicious usages. For example, when a large number of requests is originating from a single user the algorithm can detect this abnormal activity and can notify the admin regarding a potential attack.

This study discussed the train dataset noises and how dataset tuning improves the classifier model accuracy. Restricting the access of anomaly requests from various sources is also discussed.

The proposed algorithm applied to Quadratic Discriminant Analysis and Linear Discriminant Analysis models. It predicts anomalies with better precision, accuracy and F1 score.

II. RELATED WORK

In computer science, anomaly detection refers to the techniques of finding specific data points that do not conform to the normal distribution of the data set. Companies from different sectors including manufacturing, automotive, healthcare, lodging, travelling, fashion, food, and logistics are investing a lot of resources in collecting big data and exploring the hidden anomalous patterns in them to facilitate their customers. In most of the cases, the collected data are streaming time series data and due to their intrinsic characteristics of periodicity, trend, seasonality, and irregularity, it is a challenging problem to detect point anomalies precisely in them (Mohsin Munir 2018).

Anomaly-based intrusion detection systems (IDSs) have been deployed to monitor network activity and to protect systems and the Internet of Things (IoT) devices from attacks (or intrusions). The problem with these systems is that they generate a huge amount of inappropriate false alarms whenever abnormal activities are detected and they are not too flexible for a complex environment. The high-level rate of the generated false alarms reduces the performance of IDS against cyber-attacks and makes the tasks of the security analyst particularly difficult and the management of intrusion detection process computation all expensive (Wajdi Alhakami 2019).

Naive Bayes is a simple technique for classification. Naive Bayes model could be used without accepting Bayesian probability or using any Bayesian methods. An analysis of the Bayesian classification problem showed that there are sound theoretical reasons behind the apparently implausible efficacy of types of classifiers. An advantage of Naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification. The fundamental property of Naive Bayes is that it works on discrete value. If attributes values are continuous it is recommended to use Gaussian Naive Bayes classifier (Shikha Agarwal 2019).

Machine learning methods have been widely used in the intrusion detection field. Classification, clusterisation, Markov chains have been studied extensively on classical systems. In modern systems, there are many CPUs and they are shared amongst software provided by the kernel scheduler. If there is more than usual demand for CPU resources, the tasks create an order and are in standby mode for the processing. Standby regime slows down the execution time of tasks, which results in the reduction of performance metrics. To improve performance metrics CPU usage needs to be analysed. In most cases, CPU usage is analysed in terms of process, flow or task. Another metric to be analysed to improve the performance metrics is the CPU utilisation. CPU utilisation is measured in time when the CPU is engaged in the processing interval of the task and shown in the percentage. Memory access attempts also cause high CPU usage. When the i/o attempts are made, the CPU interrupts to work and waits for the process to complete (Rasim M. Alguliyev 2019).

Quadratic discriminant analysis (QDA) is a widely used classification technique that generalizes the linear discriminant analysis (LDA) classifier to the case of distinct covariance matrices among classes. For the QDA classifier to yield high classification performance, an accurate estimation of the covariance matrices is required. Such a task becomes all the more challenging in high dimensional settings, wherein the number of observations is comparable with the feature dimension (Houssem Sifaou 2020).

The performance of LDA-based classifiers depends heavily on accurate estimation of the class statistics, namely, the sample covariance matrix and class mean vectors. These statistics can be estimated with fairly high accuracy when the number of available samples is large compared to the data dimensionality. In practical high-dimensional data settings, the challenge is to cope with a limited number of available samples. In this case, the sample covariance estimates become highly perturbed and ill-conditioned resulting in severe performance degradation. In some practical situations, it occurs that the test data deviates from the training data model. For example, the training data and the test data might represent measurements obtained from non-identical devices. In such a case, the value of the regularisation parameter computed during the training phase may no longer be adequate, let alone be optimal (Alam Zaib 2021).

III. PRELIMINARIES

JSON parameters are most commonly used to send and receive data in web services. JSON allows the transfer of complex data structures efficiently, which enables the transfer of large amounts of data with minimal effort and resources.

On analyzing the above survey, we find that around 48.58% of the webservice requests come from valid users and the remaining 51.42% are found to be anomalous. Most of the requests come from invalid users, to protect against these potentially harmful requests. This proposed framework has been designed to detect and identify suspicious access patterns and automatically prevent unwanted web requests. It works by continuously monitoring the user activities and denies access for any anomalous request. It also provides a better and easy method to update rules and security policies to ensure the safety of the e-commerce webservice. It also has the option to customize security settings and fine-tune control lists and restrict user activities, etc. This paper discusses a secure environment for operating an ecommerce business.

The e-commerce dataset used to train a machine learning model to find anomalies. The dataset was prepared from API requests from various sources, ranging from 50 to 200. Each API had different parameters. Our proposed method was applied before training the Quadratic Discriminant Analysis and Linear Discriminant Analysis models. The below API retrieves product details.

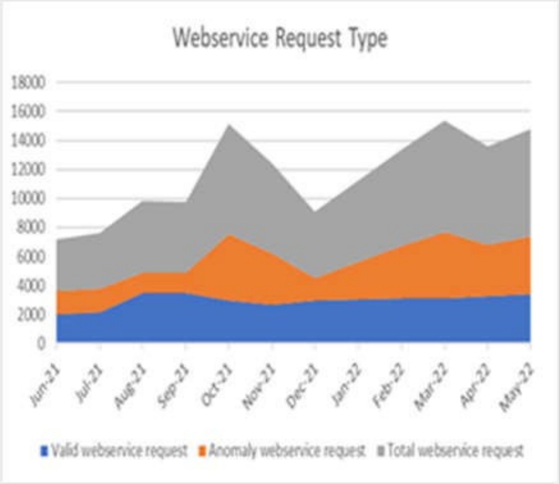

Figure 1 Chart for Month wise Count of Webservices Request Type

Dataset for get product details

```
{
        "Time": "2022-12-06 00:00",
        "API ": "resolve_product_detail",
        "data": [{
                "user_id": "retsSDWJwdw",
                "product": "Dolo 700",
                "slug": "dolo-700mg"
        }],
        "ID": “frRFEwdswwd”,
        "IP ": "172.134.04.99",
        "Country": "India",
        "OperatingSystem": "Linux"
}
```

3.1 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis is derived from the linear discriminant analysis. QDA classifies the result from two or more groups of dataset. It uses a quadratic function to classify the categories. QDA observes the difference of mean and covariance of each category. QDA provides better accuracy than LDA, with different covariance in the class.

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

The Classification rule is

$$\hat{G}(x) = \arg\max_k \delta_k(x)$$

QDA works effectively to classify the boundaries of non-linear classes.

3.2 Linear Discriminant Analysis

Linear Discriminant Analysis uses the technique of data reduction. LDA eliminates redundant data from the dataset and reduces the dimensionality of the dataset. The reduced dimension dataset enhances the between-class variance.

$$P_{lda} = \arg\max_P \frac{\left|P^T S_b P\right|}{\left|P^T S_w P\right|}$$

IV. THE PROPOSED APPROACH

The proposed method tunes the business dataset for achieving better accuracy, prediction, and F1 score. The proposed method improves the quality of the dataset. The outcome of this work gives better Quadratic Discriminant Analysis and Linear Discriminant Analysis results. This study proposes the following method to improve the dataset quality.

- Accept requests only from the e-commerce service area.
- Provide distinct user id to every guest user.
- Replace null user id by default user id.
- Replace null geo location by default geo location.

Algorithm 1

Training dataset algorithm

```
# Prepare dataset
D = {d1, d2, d3 ... dn}
# Accept requests from the e-commerce service area
G = [x for x in D if x[geo_location] = country]
# Provide every guest user with a distinct user id
I = [x for x in G if x[user_id] is guest]
x[user_id] = unique value
# Replace null user id by default user id
N = [x for x in G if x[user_id] is null]
x[user_id] = default value
# Replace null geo location by default geo location
L = [x for x in G if x[geo_location] is null]
x[geo_location] = default value
# Create a dataset with the above criteria
$DS = [ I \cap N \cap L]$
#Splitting the dataset
$DS = \{ x \in DS \mid$ 80% as training dataset and 20% as testing dataset $\}$
#Building the LDA model
LDA = build_Ldamodel(DS)
#Building the QDA model
QDA = build_Qdamodel(DS)
#Predicting the anomaly by LDA
LDA_prediction = predict_anomaly(LDA, DS)
#Predicting the anomaly by QDA
QDA_prediction = predict_anomaly(QDA, DS)
#Evaluate the result
Compare performance of LDA and QDA with Proposed dataset
```

4.1 Requests from countries outside of the E-Commerce service area is flagged

One easy but effective technique used to reduce unwanted traffic in an e-commerce platform is by restricting access to the requests that originate from outside its geographic serviceable area. This can ensure the safety of the platform from cyber attacks from other countries and also helps in reducing the traffic load of the website. This is typically done by identifying the origin of the IP address of the request and denying access to such a request that originates from outside the serviceable geo-location. In our case, all requests originating from outside India are restricted access, as our e-commerce service area is only in India.

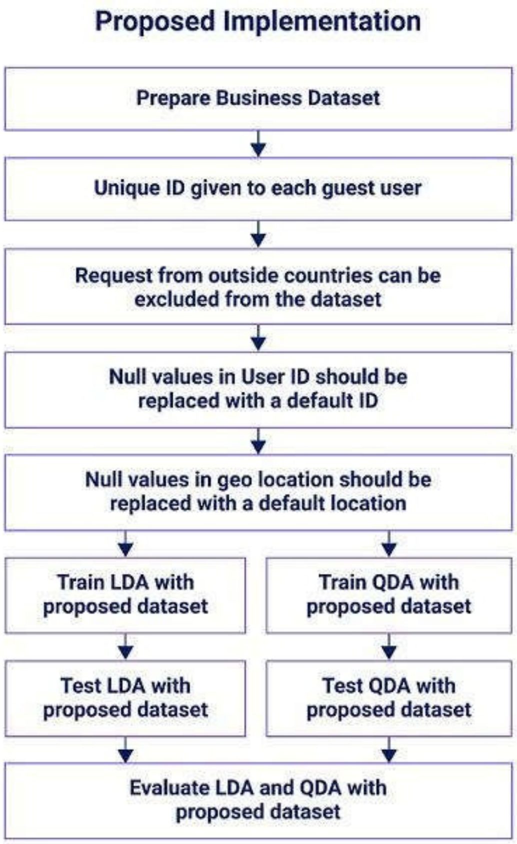
Proposed Implementation
Prepare Business Dataset
Unique ID given to each guest user
Request from outside countries can be excluded from the dataset
Null values in User ID should be replaced with a default ID
Null values in geo location should be replaced with a default location
Train LDA with proposed dataset
Train QDA with proposed dataset
Test LDA with proposed dataset
Test QDA with proposed dataset
Evaluate LDA and QDA with proposed dataset

Figure 2 Proposed Algorithm Implementation

4.2 Provide every guest user with a distinct user id

An e-commerce website receives different types of requests from different origins, and requests from users are among them. It may be from a registered user or a guest user, guest users are the types of users who are accessing the e-commerce platform but have not registered with it. They can access certain parts of the website, but some parts are restricted for them. It is important to track these users and their activities on the website. It is also important to include these users' data with the registered user's data so that the user behavior in the platform can be better understood. To track the guest user behaviors, they need to be identified individually. This can be achieved by providing the guest users with unique ids when they visit the first-time, which is known as the guest user id. A unique guest user id is provided to each guest user so that all the guest user activities can be tracked.

4.3 Replace null user id by default user-id

Null user IDs are the data in the dataset that does not have an ID for the user but all other information is present. This typically happens when a user id has not been assigned to a user or if the user id is lost when clearing the cache of the user's device. In any case, if we omit these data from the dataset, any analysis done on the dataset may become unreliable and inaccurate. So, to include these data in the analysis we are providing these null user ids with the default user id. This enables the null user ids to be included in the dataset and return, it helps in including all the user records and behaviors are included in the dataset. Also, this ensures null user ids get valid ids to be included in tracking, reporting, and other analysis. Adding the default user id to the null user id helps a lot in increasing the accuracy and comprehensiveness of the dataset.

4.4 Replace null geolocation with default geolocation

Null geo-location is when user data in a dataset has no geo-location in it. This typically happens when a location entered by a user is wrong or not identified or if the GPS coordinates of the user are not available or if the IP address cannot be geo-located. The null geo location has to be changed to the default geo-location to include the particular user data to be included in the dataset. By replacing the null geo-location with the default geo-location, we are assigning an approximate location for the user for whom the true geo-location cannot be identified. This technique enables the system to complete the user data and helps in increasing the accuracy when analyzing the dataset. In our case, we use the default location as India.

V. EXPERIMENTAL RESULT

The proposed dataset is tested with Quadratic Discriminant Analysis and Linear Discriminant Analysis classifiers. The experimental result is derived in the below section.

5.1 Quadratic Discriminant Analysis with proposed algorithm

The data shows the Confusion Matrix of Quadratic Discriminant Analysis.

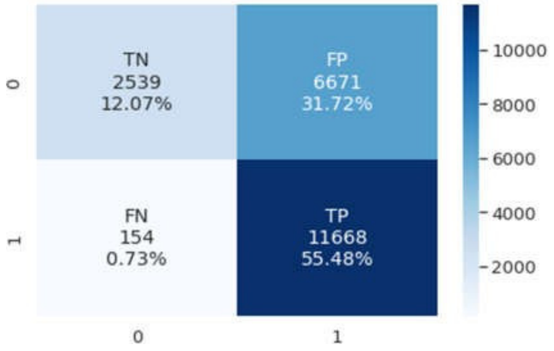

Figure 3 Confusion Matrix of Quadratic Discriminant Analysis with proposed algorithm

The proposed algorithm performance metrics are calculated from True Positive, True Negative, False Positive and False Negative.

True Positive (TP) = 11668

True Negative (TN) = 2539

False Positive (FP) = 6671

False Negative (FN) = 154

5.1 Accuracy

The accuracy is the percentage of correct predictions out of the total predictions.

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

$$Accuracy = \frac{(11668 + 2539)}{(11668 + 2539 + 6671 + 154)}$$

Accuracy = 0.6754944846

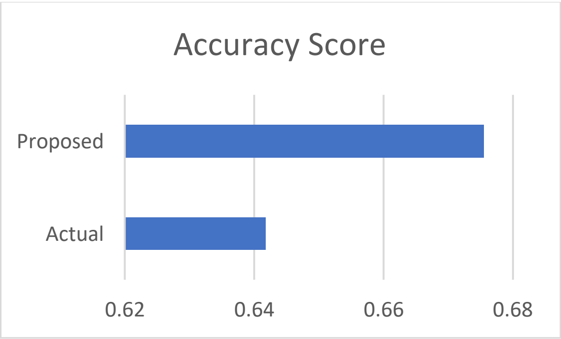

Figure 4 Accuracy score of QDA

The Accuracy score has improved by 5.24%.

5.2 Precision

The precision is the metric that defines the model's ability to predict the correct number of yes outcomes out of the total yes outcomes predicted.

$$Precisiok = \frac{TP}{(TP + FP)}$$

$$Precisiok = \frac{11668}{(11668 + 6671)}$$

Precision = 0.6362397077

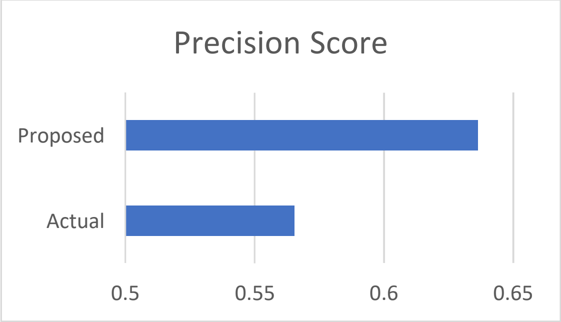

Figure 5 Precision score of QDA

The Precision score has improved by 12.53%.

5.3 Recall

The recall compares the model's number of correct yes outcomes predicted with the total number of actual yes outcomes.

$$Recall = \frac{TP}{(TP + FN)}$$

$$Recal = \frac{11668}{(11668 + 154)} = 0.9869734394$$

5.4 F1 Score

F1 Score combines both precision and recall scores for fool proof evaluation.

$$F1\,Score = \frac{2\ x\ (precisiok\ score\ x\ recall\ score)}{(precisiok\ score + recall\ score)}$$

$$F1\ Score = \frac{2\,x\,(\,0.6362397077\,x\ 0.9869734394)}{(0.6362397077 +\ 0.9869734394)}$$

F1 Score = 0.7737143994

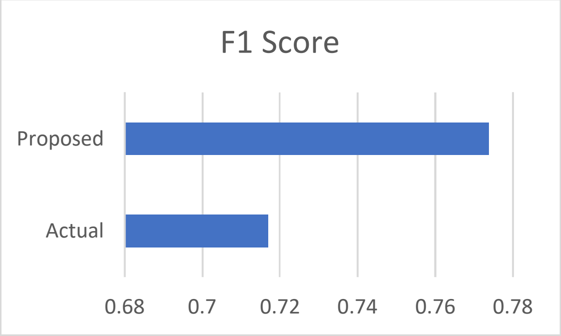

Figure 6 F1 Score of QDA

The F1 score has improved by 7.91%.

5.5 Miss Rate

The miss rate measures the incorrect predictions. The positive result that is predicted as negative is known as a false negative.

$$Miss\ Rate = \frac{FN}{FN + TP}$$

$$Miss\ Rate = \frac{154}{154 + 11668}$$

Miss Rate = 0.01302656065

5.6 Fall-out

The fallout rate measures the incorrect predictions. The negative results predicted as positive are known as a false positive.

$$Fall - out\ = \frac{FP}{FP + TN}$$

$$Fall - out = \frac{6671}{6671 + 2539}$$

Fall-out = 0.7243213898

5.7 Specificity

Specificity is the metric used to find the true negatives; it is also referred to as the True Negative Rate.

$$Specificity = \frac{TN}{TN + FP}$$

$$Specificity\ = \frac{2539}{2539 + 6671}$$

Specificity = 0.2756786102

5.8 Mean Absolute Error

The Mean absolute error value is 32.45% for the proposed algorithm. It is reduced by 3.4%. Quadratic Discriminant Analysis is a classifier model tested with the proposed algorithm. The results, shown in the below table, were compared using accuracy score, F1 score, precision score, recall, miss rate, fall-out, specificity and mean absolute error. The results demonstrate that the proposed algorithm is improving the anomaly detection performance.

Table 1 Result Comparison of Quadratic Discriminant Analysis with proposed algorithm

Metrics	Quadratic Discriminant Analysis	Proposed algorithm
Accuracy Score	64.18%	67.55%
F1 Score	71.69%	77.37%
Precision Score	56.54%	63.62%
Recall	97.96%	98.70%
Miss Rate	2.04%	1.30%
Fall-out	64.95%	72.43%
Specificity	35.05%	27.57%
Mean Absolute Error	35.82%	32.45%

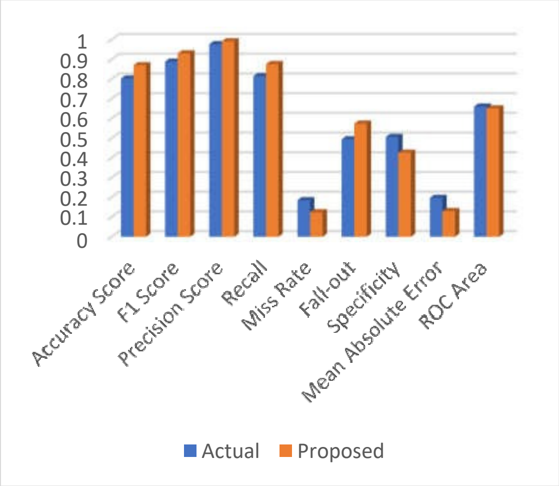

Figure 7 Performance Metrics of Quadratic Discriminant Analysis with proposed algorithm

5.2 Linear Discriminant Analysis with proposed algorithm

The proposed algorithm performance metrics are calculated from True Positive, True Negative, False Positive and False Negative.

True Positive (TP) = 18187
True Negative (TN) = 113
False Positive (FP) = 152
False Negative (FN) = 2580

The data shows Confusion Matrix of Linear Discriminant Analysis.

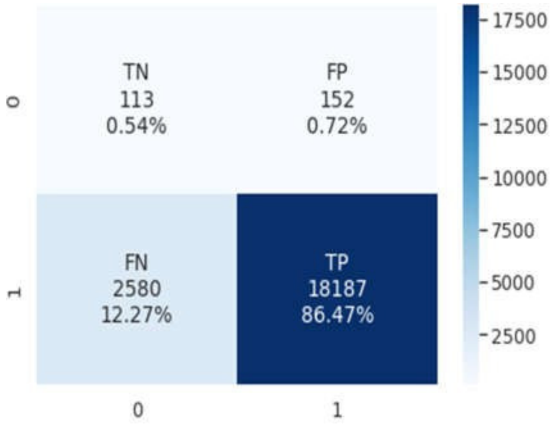

Figure 8 Confusion Matrix of Linear Discriminant Analysis with proposed algorithm

5.1 Accuracy

The accuracy is the percentage of correct predictions out of the total predictions.

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

$$Accuracy = \frac{(18187 + 113)}{(18187 + 113 + 152 + 2580)}$$

Accuracy = 0.870103

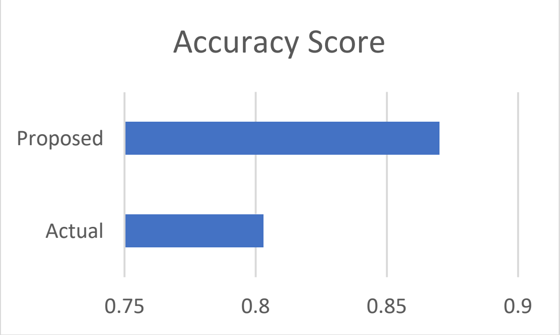

Figure 9 Accuracy score of LDA

The Accuracy score has improved by 8.37%.

5.2 Precision

The precision is the metric that defines the model's ability to predict the correct number of yes outcomes out of the total yes outcomes predicted.

$$Precisiok = \frac{TP}{(TP + FP)}$$

$$Precisiok = \frac{18187}{(18187 + 152)} = 0.991712$$

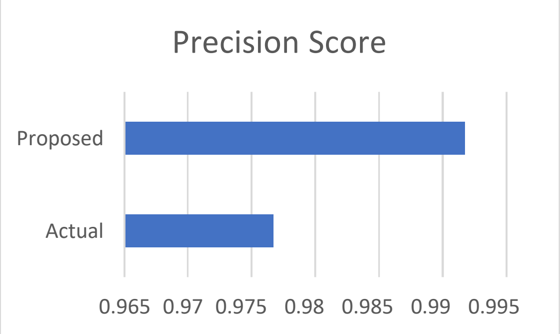

Figure 10 Precision score of LDA

The Precision score has improved by 1.54%.

5.3 Recall

The recall compares the model's number of correct yes outcomes predicted with the total number of actual yes outcomes.

$$Recall = \frac{TP}{(TP + FN)}$$

$$Recal = \frac{18187}{(18187 + 2580)} = 0.875764434$$

5.4 F1 Score

F1 Score combines both precision and recall scores for fool proof evaluation.

$$F1\,Score = \frac{2\,x\,(precisiok\ score\ x\ recall\ score)}{(precisiok\ score + recall\ score)}$$

$$F1\,Score = \frac{2\,x\,(\,0.991712\,x\ 0.875764434)}{(0.991712 + 0.875764434)}$$

F1 Score = 0.9301385977

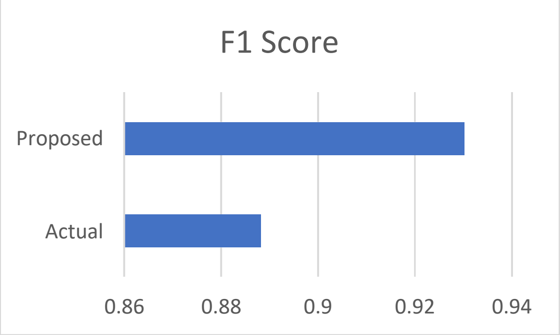

Figure 11 F1 Score of LDA

The F1 score has improved by 4.71%.

5.5 Miss Rate

The miss rate measures the incorrect predictions. The positive result that is predicted as negative is known as a false negative.

$$Miss\ Rate = \frac{FN}{FN + TP}$$

$$Miss\ Rate = \frac{2580}{2580 + 18187}$$

Miss Rate = 0.124235566

5.6 Fall-out

The fallout rate measures the incorrect predictions. The negative results predicted as positive are known as a false positive.

$$Fall - ou = \frac{FP}{FP + TN}$$

$$Fall - out = \frac{152}{152 + 113}$$

Fall-out = 0.5735849057

5.7 Specificity

Specificity is the metric used to find the true negatives; it is also referred to as the True Negative Rate.

$$Specificity = \frac{TN}{TN + FP}$$

$$Specificity = \frac{113}{113 + 152}$$

Specificity = 0.4264150943

5.8 Mean Absolute Error

The Mean absolute error value is 12.99% for the proposed algorithm. It is reduced by 6.73%.

Table 2 Result Comparison of Linear Discriminant Analysis with proposed algorithm

Metrics	Linear Discriminant Analysis	Proposed algorithm
Accuracy Score	80.28%	87.01%
F1 Score	88.83%	93.01%
Precision Score	97.67%	99.17%
Recall	81.45%	87.58%
Miss Rate	18.55%	12.42%
Fall-out	49.39%	57.36%
Specificity	50.61%	42.64%
Mean Absolute Error	19.72%	12.99%

Linear Discriminant Analysis is a classifier model tested with the proposed algorithm. The results, shown in the above table, were compared using accuracy score, F1 score, precision score, recall, miss rate, fall-out, specificity and mean absolute error. The results demonstrate that the proposed algorithm is improving the anomaly detection performance.

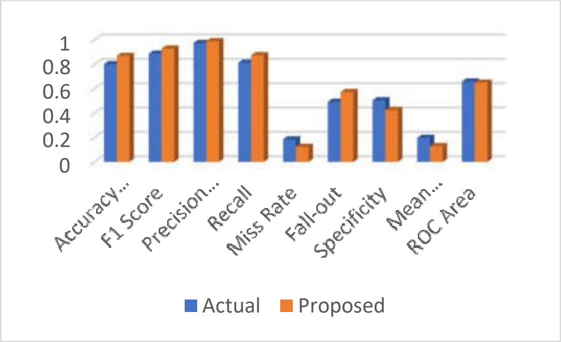

Figure 12 Performance Metrics of Linear Discriminant Analysis with proposed algorithm

VI CONCLUSION

The proposed algorithm tested with Quadratic Discriminant Analysis and Linear Discriminant Analysis models. It improves Accuracy score, Precision score and F1 score.

Quadratic Discriminant Analysis with proposed algorithm. The Accuracy score has improved by 5.24%. The Precision score has improved by 12.53%. The F1 score has improved by 7.91%.

Linear Discriminant Analysis with proposed algorithm. The Accuracy score has improved by 8.37%. The Precision score has improved by 1.54%. The F1 score has improved by 4.71%.

VII FUTURE WORK

The proposed algorithm can be used with different classifiers. They are Gaussian Naive Bayes, Logistic Regression, MLP Classifier, Adaboost Classifier, Random Forest Classifier, Decision Tree Classifier, and K-Nearest Neighbors. Train the models with different domain datasets. The wide variety of training datasets predicts with better accuracy. The dataset can be retrieved from different seasons and different geographical location people data.

REFERENCES

[1] Bayu Adhi Tama, Lewis Nkenyereye, S. M. Riazul Islam, and Kyung-Sup Kwak (2020).An Enhanced Anomaly Detection in Web Traffic Using a Stack of Classifier Ensemble. IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.2969428, pages 24120 – 24134.

[2] Wajdi Alhakami, Abdullah Alharbi, Sami Bourouis, Roobaea Alroobaea, and Nizar Bouguila (2019). Network Anomaly Intrusion Detection Using a Nonparametric Bayesian Approach and Feature Selection. IEEE Access, *Digital Object Identifier 10.1109/ACCESS.2019.2912115, pages* 52181 – 52190.

[3] Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, Jie Chen, Zhaogang Wang, Honglin Qiao (2018). Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications, Cornell University.

[4] Ya Su, Youjian Zhao,Chenhao Niu,Rong Liu,Wei Sun, Dan Pei (2019). Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining

[5] Venkat N. Gudivada, Amy Apon, and Junhua Ding (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. International Journal on Advances in Software 10.1, pages 1 - 20.

[6] Alam Zaib, Tarig Ballal, Shahid Khattak, and Tareq Y. Al-Naffouri (2021), A Doubly

Regularized Linear Discriminant Analysis Classifier With Automatic Parameter Selection. IEEE Access, Digital Object Identifier 10.1109/ACCESS.2021.3068611, pages 51343 – 51354.

[7] Houssem Sifaou, Abla Kammoun, and Mohamed-Slim Alouini (2020), High-Dimensional Quadratic Discriminant Analysis Under Spiked Covariance Model. IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.3004812, pages 117313 – 117323.

[8] Chen Zhuang, Hongbo Zhao, Chao Sun, and Wenquan Feng (2019), Detection and Classification of GNSS Signal Distortions Based on Quadratic Discriminant Analysis. IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.2965617, pages 25221 – 25236.

[9] V. Sankar, Dr. G. Zayaraz(2020). JSON and XML data security performance of Encrypted File Process in Web Services. Published in International Journal of Current Science (IJCSPUB), Volume 10, Issue 4.

[10] Ravi Chandra Jammalamadaka, Sharad Mehrotra (2006). Querying Encrypted XML Documents. Published International Database Engineering and Applications Symposium (IDEAS'06)

[11] RA. K. Saravanaguru, George Abraham, Krishnakumar Venkatasubramanian, Kiransinh Borasia (2013). Securing Web Services Using XML Signature and XML Encryption. School of Computer Science and Engineering,VIT University, Vellore, India.

[12] Hoi Ting Poon and Ali Miri (2015). Computation and Search over Encrypted XML Documents. Department of Computer Science Ryerson University Toronto, Ontario, Canada, 978-1-4673-7278-7/15

[13] Gu Yue-sheng, Ye Meng-tao, Gan Yong (2010). Web Services Security Based on XML Signature and XML Encryption. Journal of networks, Vol. 5, No. 9.

[14] Nithin N and Anupkumar M Bongale (2012). XBMRSA:A New XML Encryption Algorithm. Proceedings of Information and Communication Technologies. World Congress,pp 567-571,2012.

[15] Nithin N and Harshitha.K.S (2014). Analysis of Symmetric algorithm for XML document security. International Journal of Innovations in Engineering and Technology (IJIET), Vol. 3 Issue 4.

[16] Aamer Nadeem (2005). A Performance Comparison of Data Encryption Algorithms. IEEE.

[17] Ravi Varma1, Dr. G. Venkat Rami Reddy (2014). Schema Based Parallel XML Parser: A Fast XML Parser Designed forLarge XML Files. International Journal of Computer Science and Mobile Computing. Vol.3 Issue.8.

[18] Shikha Agarwal, Balmukumd Jha, Tisu Kumar, Manish Kumar, Prabhat Ranjan(2019). Hybrid of Naive Bayes and Gaussian Naive Bayes for Classification: A Map Reduce Approach. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-6S3.

[19] Rasim M. Alguliyev, Ramiz M. Aliguliyev and Fargana Jabbar Abdullayeva (2019). Hybridisation of classifiers for anomaly detection in big data. International Journal of Big Data Intelligence. 10.1504/IJBDI.2019.10018528