ELSEVIER

Scopus

doi

Journal of Vibration Engineering

ISSN:1004-4523

Registered

- SCOPUS
- GOOGLE SCHOLAR
- DIGITAL OBJECT IDENTIFIER (DOI)
- IMPACT FACTOR 6.1

Our Website
www.jove.science

Enhancing a Dataset to Improve Anomaly Detection UsingMachineLearning

V.Sankar,ResearchScholar,DepartmentofComputerScience,BharathiarUniversity,Coimbatore, TamilNadu,India
Dr.G.Zayaraz,Professor,DepartmentofCSE,PondicherryEngineeringCollege,Pondicherry

ABSTRACTThispaperdiscussesthedatasecurityof customer andbusinessdata.Thebusinesswebsitereceivesrequestsfromvarioussources such as customers, crawlers, bots, andhackers. The elimination of anomalies improvesthe effective usage of hardware, software, andnetworkbandwidth.Amachinelearningtechnique is used to identify anomalies fromeachAPIrequest.

Thequalityofthedatasetisimportantforbetteraccuracy. This paper proposes to improve thequality of the training dataset using QuadraticDiscriminantAnalysisandLinearDiscriminantAnalysismodels.Theproposedmethodis expectedtoyieldbetteraccuracy,precision,andF1-score.

Keywords – Quadratic Discriminant AnalysisandLinearDiscriminantAnalysis,Web service security, Anomaly detection, MachineLearning.

I. INTRODUCTION

E-Commerce security ensures customer dataarestoredsafelyandalltransactionsarecarriedout without any compromise in data security.Cyberattacksone-commercemaycomeinmanyformsandproperpreventivemeasuresareto be takentoensure data protection. Somecommontypesofattacksarephishing,monetary theft, credit card fraud, hacking, andmisusingofintellectualproperty.

Anye-commerceplatformensuresproperprotection against these threats and prioritizesecuritymeasuresforitscustomers.The machine learning mechanism is used to restrictanomalyaccess.

Analyzing datasets using machine learningcan help in processing huge datasets and detectpatternsthatcouldbeusedtoisolateattacksandmalicious usages. For example, when a largenumber of requests is originating from a singleuserthealgorithmcandetectthisabnormalactivity and can notify the admin regarding apotentialattack.

This study discussed the train dataset noisesand how dataset tuning improves the classifiermodelaccuracy.Restrictingtheaccessof anomaly requests from various sources is alsodiscussed.

TheproposedalgorithmappliedtoQuadraticDiscriminantAnalysisandLinearDiscriminantAnalysismodels.Itpredictsanomalieswithbetterprecision,accuracyandF1score.

II. RELATEDWORK

Incomputerscience,anomalydetectionrefers to the techniques of finding specific datapointsthatdonotconformtothenormaldistribution of the data set. Companies fromdifferentsectorsincludingmanufacturing,automotive,healthcare,lodging,travelling,fashion,food,andlogisticsareinvestingalotofresources in collecting big data and exploringthehiddenanomalouspatternsinthemto facilitate their customers. In most of the cases,thecollecteddataarestreamingtimeseriesdataandduetotheirintrinsiccharacteristicsofperiodicity, trend, seasonality, and irregularity,itisachallengingproblemtodetectpointanomaliespreciselyinthem(MohsinMunir2018).

Anomaly-based intrusion detection systems(IDSs)havebeendeployedtomonitornetwork

activity and to protect systems and the InternetofThings(IoT)devicesfromattacks(orintrusions).The problem with these systems isthattheygenerateahugeamountofinappropriate false alarms whenever abnormalactivitiesaredetectedandtheyarenottoo flexible for a complex environment. The high-level rate of the generated false alarms reducesthe performance of IDS against cyber-attacksandmakesthetasksofthesecurityanalystparticularlydifficultandthemanagementofintrusiondetectionprocesscomputationallexpensive(WajdiAlhakami2019).

NaiveBayesisasimpletechniqueforclassification.NaiveBayesmodelcouldbeused without accepting Bayesian probability orusinganyBayesianmethods.AnanalysisoftheBayesianclassificationproblemshowedthatthere are sound theoretical reasons behind theapparentlyimplausibleefficacyoftypesofclassifiers.AnadvantageofNaiveBayesisthatit only requires a small number of training datatoestimatetheparametersnecessaryforclassification.ThefundamentalpropertyofNaive Bayes is that it works on discrete value.Ifattributesvaluesarecontinuousitisrecommendedtouse GaussianNaiveBayesclassifier(ShikhaAgarwal2019).

Machinelearningmethodshavebeenwidelyusedintheintrusiondetectionfield.Classification,clusterisation,Markovchainshavebeenstudiedextensivelyonclassicalsystems. In modern systems, there are manyCPUs and they are shared amongst softwareprovided by the kernel scheduler. If there ismorethanusualdemandforCPUresources,thetasks create an order and are in standby modefortheprocessing.Standbyregimeslowsdownthe executiontimeoftasks,whichresultsintherduction of performance metrics. To improveperformance metrics CPU usage needs to beanalysed.Inmostcases,CPUusageisanalysedintermsofprocess,flowortask.Anothermetrictobeanalysedtoimprovetheperformancemetricsisthe CPUutilisation.CPU utilisation is measured in time when theCPUisengagedintheprocessingintervaloftheeduction taskandshowninthepercentage.Memoryaccess attempts also cause high CPU usage.Whenthei/oattemptsaremade,theCPUinterrupts to work and waits for the process tocomplete(RasimM.Alguliyev2019).

Quadratic discriminant analysis (QDA) is awidelyusedclassificationtechniquethatgeneralizesthelineardiscriminantanalysis(LDA)classifiertothecaseofdistinctcovariancematricesamongclasses.FortheQDAclassifiertoyieldhighclassificationperformance,anaccurateestimationofthecovariance matrices is required. Such a taskbecomesallthemorechallenginginhighdimensional settings, wherein the number ofobservationsiscomparablewiththefeaturedimension(HoussemSifaou2020).

The performance of LDA-based classifiersdepends heavily on accurate estimation of theclass statistics, namely, the sample covariancematrix and class mean vectors. These statisticscanbeestimatedwithfairlyhighaccuracy whenthenumberofavailablesamplesislargecomparedtothedatadimensionality.Inpracticalhigh-dimensionaldatasettings,thechallenge is to cope with a limited number ofavailablesamples.Inthiscase,thesamplecovariance estimates become highly perturbedandill-conditionedresultinginsevereperformancedegradation.Insomepracticalsituations, it occurs that the test data deviatesfrom the training data model. For example, thetraining data and the test data might representmeasurementsobtainedfromnon-identicaldevices.Insuchacase,thevalueofthereguларisation parameter computed during thetraining phase may no longer be adequate, letalonebeoptimal(AlamZaib2021).

III. PRELIMINARIES

JSON parameters are most commonly usedtosendandreceivedatainwebservices.JSONallows the transfer of complex data structuresefficiently, which enables the transfer of largeamountsofdatawithminimaleffortandresources.

On analyzing the above survey, we find thataround48.58%ofthewebservicerequestsco mefromvalidusersandtheremaining51.42%aref oundtobeanomalous.Mostoftherequests come from invalid users, to protectagainstthesepotentiallyharmfulrequests. Thisproposedframeworkhasbeendesignedtodet ect and identify suspicious access patternsandautomaticallypreventunwantedweb requests. It works by continuously monitoringthe user activities and denies access for anyanomalousrequest.Italsoprovidesabetteran deasymethodtoupdaterulesandsecuritypoliciest oensurethesafetyofthee-commercewebservice. It also has the option to customizesecurity settings and fine-tune control lists andrestrictuseractivities,etc.Thispaperdiscusse sasecureenvironmentforoperatinganecommerc ebusiness.

Thee-commercedatasetusedtotrainamachinelearning modeltofindanomalies.Thedataset was prepared from API requests fromvarious sources, ranging from 50 to 200. EachAPI had different parameters. Our proposedmethodwasappliedbeforetrainingthe QuadraticDiscriminant Analysis and LinearDiscriminantAnalysismodels.Thebelow APIretrievesproductdetails.

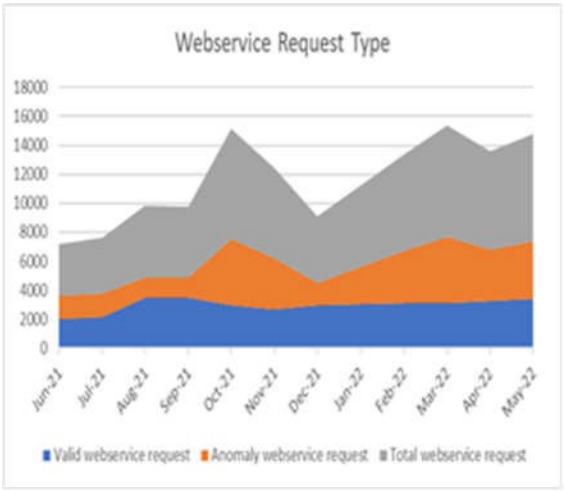

Figure 1 Chart for Month wise Count ofWebservicesRequestType

Datasetforgetproductdetails


```
{
        "Time":"2022-12-0600:00",
        "API":"resolve_product_detail","
        data":[{
                "user_id":
                "retsSDWJwdw","product":
                "Dolo 700","slug":"dolo-
                700mg"
        }],
        "ID":“frRFEwdswwd”,
        "IP":"172.134.04.99",
        "Country":
        "India","OperatingSystem":"Linu
        x"
}
```


3.1 QuadraticDiscriminantAnalysis

QuadraticDiscriminantAnalysisisderi ved from the linear discriminant analysis.QDA classifies the result from two or moregroups of dataset. It uses a quadratic functionto classify the categories. QDA observes thedifference of mean and covariance of eachcategory. QDA provides better accuracy thanLDA,withdifferentcovarianceintheclass.

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

TheClassificationruleis

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

QDA works effectively to classify theboundariesofnon-linearclasses.

3.2 LinearDiscriminantAnalysis

LinearDiscriminantAnalysisusesthetechni que of data reduction. LDA eliminatesredundant data from the dataset and reducesthedimensionalityofthedataset.Thered uceddimensiondatasetenhancesthebetween-classvariance.

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

IV. THEPROPOSEDAPPROACH

The proposed method tunes the businessdatasetforachievingbetteraccuracy,prediction,andF1score.Theproposedmethod improves the quality of the dataset.TheoutcomeofthisworkgivesbetterQuadratic Discriminant Analysis and LinearDiscriminantAnalysisresults.Thisstudy proposesthefollowingmethodtoimprovethedatasetquality.

- Accept requests only from the e-commerceservicearea.
- Providedistinctuseridtoeveryguestuser.
- Replacenulluseridbydefaultuserid.
- Replacenullgeolocationbydefaultgeolocation.

Algorithm1

Trainingdatasetalgorithm

#Preparedataset
D={d1,d2,d3...dn}
#Acceptrequestsfromthee-commerceservicearea
G=[xforxinDifx[geo_location]=country]
#Provideeveryguestuserwithadistinctuserid
I=[xforxinGifx[user_id]isguest]x[user_id]=uniquevalue
#ReplacenulluseridbydefaultuseridN = [x for x in G if x[user_id] is null]x[user_id]=defaultvalue
#Replacenullgeolocationbydefaultgeolocation
L=[xforxinGifx[geo_location]isnull]x[geo_location]=defaultvalue
#Createadatasetwiththeabovecriteria
DS=[I∩N∩L]
#Splittingthedataset
DS={x∈DS|80%astrainingdatasetand20%astestingdataset}
#Building the LDA modelLDA=build_Ldamodel (DS)
#Building the QDA modelQDA=build_Qdamodel (DS)
#Predicting the anomaly by LDALDA_prediction = predict_anomaly(LDA,DS)
#Predicting the anomaly by QDAQDA_prediction= predict_anomaly(QDA,DS)
#Evaluatetheresult
CompareperformanceofLDAandQDAwithProposeddataset

4.1 Requests from countries outside of theE-Commerceserviceareaisflagged

Oneeasybuteffectivetechniqueusedtoreduce unwantedtrafficinane-commerceplatform is by restricting access to the requeststhatoriginatefromoutsideitsgeographic serviceable area. This can ensure the safety oftheplatformfromcyberattacksfromothercountries and also helps in reducing the trafficload of the website. This is typically done byidentifying the origin of the IP address of therequest and denying access to such a requestthat originates from outside the serviceable geo-location. In our case, all requests originatingfrom outside India are restricted access, as oure-commerceserviceareaisonlyinIndia.

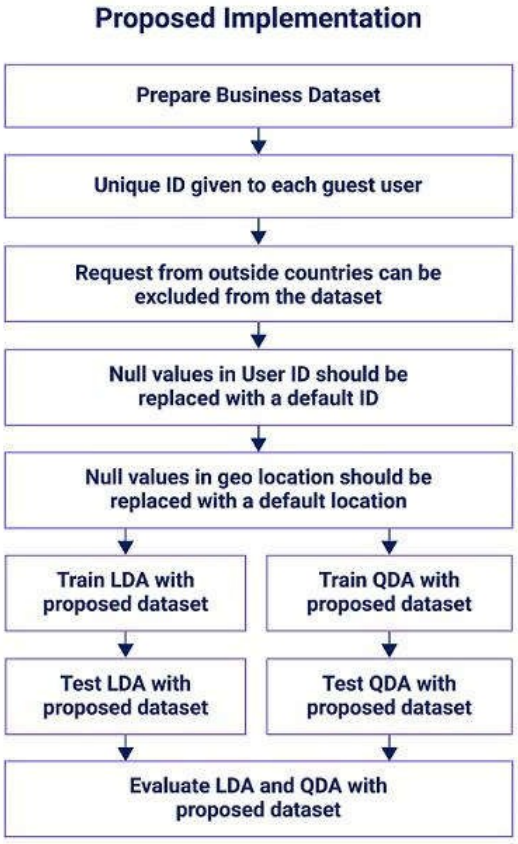

Figure 2 Proposed AlgorithmImplementation

4.2 Provide every guest user with a distinctuserid

An e-commerce website receives differenttypes of requests from different origins, andrequests from users are among them. It may befrom a registered user or a guest user, guestusers are the types of users who are accessingthee-commerceplatformbuthavenotregisteredwithit. Theycanaccesscertainpartsof the website, but some parts are restricted forthem. It is important to track these users andtheiractivitiesonthewebsite.Itisalsoimportant to include these users' data with theregistered user's data so that the user behaviorin the platform can be better understood. Totrack the guest user behaviors, they need to beidentifiedindividually.Thiscanbeachievedby providing the guest users with unique ids whenthey visit the first-time, which is known as theguestuserid.Auniqueguestuseridisprovided toeachguestusersothatalltheguestuseractivitiescanbetracked.

4.3 Replacenulluseridbydefaultuser-id

Null user IDs are the data in the dataset thatdoes not have an ID for the user but all otherinformation is present. This typically happenswhen a user id has not been assigned to a useror if the user id is lost when clearing the cacheof the user's device.In any case, if we omitthese data from the dataset, any analysis doneonthedatasetmaybecomeunreliableandinaccurate.So,toincludethesedataintheanalysis we are providing these null user idswith the default user id. This enables the nulluseridstobeincludedinthedatasetandreturn,it helps in including all the user records andbehaviors are included in the dataset. Also, thisensuresnulluseridsgetvalididstobeincludedintracking,reporting,andotheranalysis.Adding the default user id to the null user idhelpsalotinincreasingtheaccuracyandcomprehensivenessofthedataset.

4.4 Replace null geolocation with defaultgeolocation

Nullgeo-locationiswhenuserdatainadataset has no geo-location in it. This typicallyhappens when a location entered by a user iswrongornotidentifiedoriftheGPScoordinatesoftheuserarenotavailableoriftheIP address cannot be geo-located. The null geolocation has to be changed to the default geo-locationtoincludetheparticularuserdatatobeincluded in the dataset.By replacing the nullgeo-location with the default geo-location, weare assigning an approximate location for theuser for whom the true geo-location cannot beidentified.Thistechniqueenablesthesystemtocomplete the user data and helps in increasingtheaccuracywhenanalyzingthedataset.Inourcase,weusethedefaultlocationasIndia.

V. EXPERIMENTALRESULT

TheproposeddatasetistestedwithQuadraticDiscriminantAnalysisandLinearDiscriminantAnalysis classifiers. The experimental result isderivedinthebelowsection.

5.1 QuadraticDiscriminantAnalysiswithproposedalgorithm

The datashows the Confusion Matrix of QuadraticDiscriminantAnalysis.

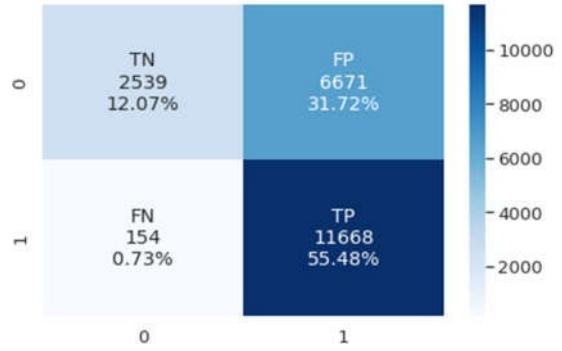

Figure3ConfusionMatrixofQuadraticDiscriminantAnalysiswithproposedalgorithm

Theproposedalgorithmperformancemetricsarecalculatedfrom TruePositive,TrueNegative,FalsePositiveandFalseNegative.

TruePositive(TP)=11668TrueNegative(TN)=2539False Positive (FP) = 6671FalseNegative(FN)=154

5.1Accuracy

The accuracy is the percentage of correctpredictionsoutofthetotalpredictions.

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

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

Accuracy=0.6754944846

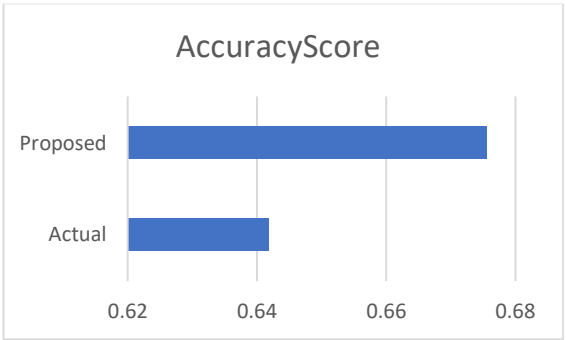

Figure4AccuracyscoreofQDA

TheAccuracyscorehasimprovedby5.24%.

5.2 Precision

The precision is the metric that defines themodel's ability to predict the correct number ofyesoutcomesoutofthetotalyesoutcomespredicted.

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

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

Precision=0.6362397077

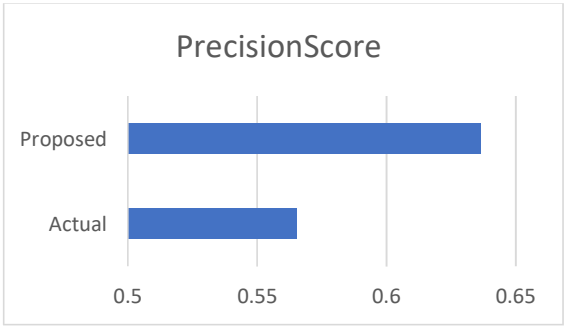

Figure5PrecisionscoreofQDA

ThePrecisionscorehasimprovedby12.53%.

5.3 Recall

Therecallcomparesthemodel'snumberof correctyesoutcomespredictedwiththetotalnumberofactualyesoutcomes.

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

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

5.4 F1Score

F1 Score combines both precision and recallscoresforfoolproofevaluation.

$$F1Score = \frac{2x(precisiokscorexrecallscore)}{(precisiokscore+recallscore)}$$

$$F1Score = \frac{2x(0.6362397077x0.9869734394)}{(0.6362397077+0.9869734394)}$$

F1Score=0.7737143994

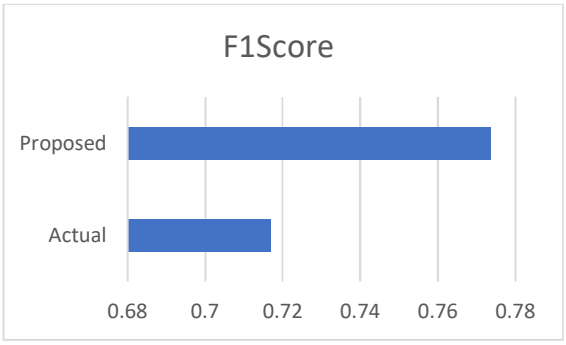

Figure6F1ScoreofQDA

TheF1scorehasimprovedby7.91%.

5.5 MissRate

Themissratemeasurestheincorrectpredictions. Thepositiveresultthatispredictedasnegativeis knownasafalsenegative.

$$MissRate = \frac{FN}{FN+TP}$$

$$MissRate = \frac{154}{154+11668}$$

MissRate=0.01302656065

5.6 Fall-out

Thefalloutratemeasurestheincorrectpredictio ns. The negative results predicted aspositiveareknownasafalsepositive.

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

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

Fall-out=0.7243213898

5.7 Specificity

Specificity is the metric used to find the truenegatives;itisalsoreferredtoastheTrueNegat iveRate.

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

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

Specificity=0.2756786102

5.8MeanAbsoluteError

The Mean absolute error value is 32.45% forthe proposed algorithm. It is reduced by 3.4%.QuadraticDiscriminantAnalysisisaclassi fiermodeltestedwiththeproposedalgorithm. The results, shown in the belowtable, were compared using accuracy score,F1 score, precision score, recall, miss rate,fall-out, specificity and mean absolute error.Theresultsdemonstratethattheproposeda lgorithmisimprovingtheanomalydetectionper formance.

Table1ResultComparisonofQuadraticDiscrimi nantAnalysiswithproposedalgorithm

Metrics	QuadraticDiscriminantAnalysis	Proposed algorithm
AccuracyScore	64.18%	67.55%
F1Score	71.69%	77.37%
PrecisionScore	56.54%	63.62%
Recall	97.96%	98.70%
MissRate	2.04%	1.30%
Fall-out	64.95%	72.43%
Specificity	35.05%	27.57%
Mean Absolute Error	35.82%	32.45%

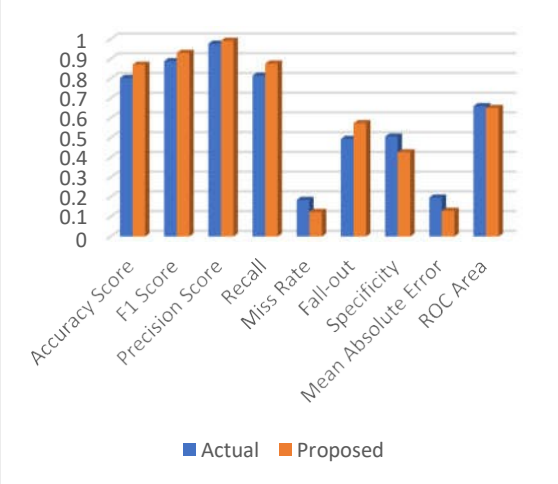

Figure 7 Performance Metrics of QuadraticDiscriminantAnalysiswithproposedalgorithm

5.2 LinearDiscriminantAnalysiswithproposedalgorithm

Theproposedalgorithmperformancemetricsarecalculatedfrom TruePositive,TrueNegative,FalsePositiveandFalseNegative.

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

ThedatashowsConfusionMatrixofLinearDiscriminantAnalysis.

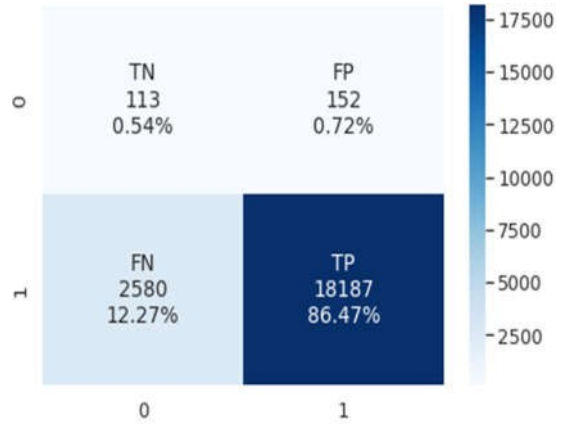

Figure 8 Confusion Matrix of LinearDiscriminantAnalysiswithproposedalgorithm

5.1 Accuracy

Theaccuracyisthepercentageofcorrectpredictionsoutofthetotalpredictions.

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

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

Accuracy=0.870103

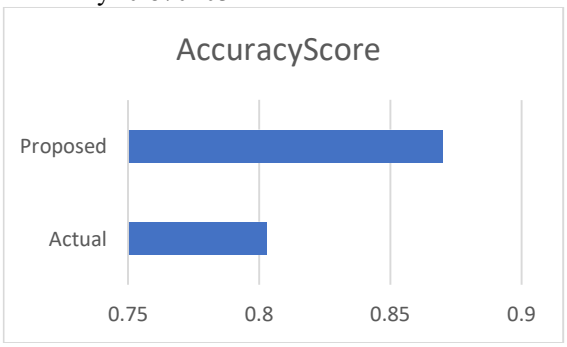

Figure9AccuracyscoreofLDA

TheAccuracyscorehasimprovedby8.37%.

5.2 Precision

The precision is the metric that defines themodel's ability to predict the correct number ofyesoutcomesoutofthetotalyesoutcomespredicted.

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

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

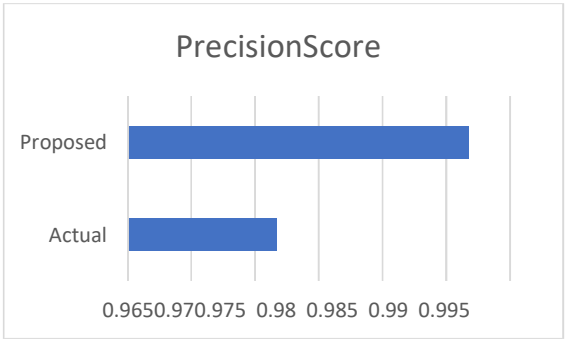

Figure10PrecisionscoreofLDA

ThePrecisionscorehasimprovedby1.54%.

5.3 Recall

The recall compares the model's number ofcorrect yes outcomes predicted with the totalnumberofactualyesoutcomes.

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

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

5.4F1Score

F1 Score combines both precision and recallscoresforfoolproofevaluation.

$$F1Score = \frac{2x(precisiokscorexrecallscore)}{(precisiokscore+recallscore)}$$

$$F1Score= \frac{2x(0.991712x0.875764434)}{(0.991712+0.875764434)}$$

F1Score=0.9301385977

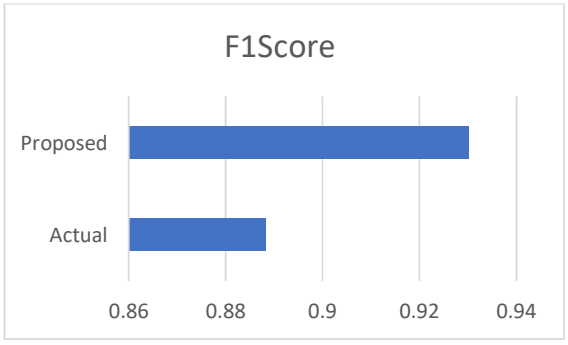

Figure11F1ScoreofLDA

TheF1scorehasimprovedby4.71%.

5.5MissRate

The miss rate measures the incorrect predictions.Thepositiveresultthatispredicteda snegativeisknownasafalsenegative.

$$MissRate= \frac{FN}{FN+TP}$$

$$MissRate= \frac{2580}{2580+18187}$$

MissRate=0.124235566

5.6Fall-out

The fallout rate measures theincorrectpredictions.Thene gativeresultspredictedas positiveareknownasafalsepositive.

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

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

Fall-out=0.5735849057

5.7Specificity

Specificity is the metric used to find the truenegatives;itisalsoreferredtoastheTrueNegat iveRate.

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

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

Specificity=0.4264150943

5.8MeanAbsoluteError

TheMeanabsoluteerrorvalueis12.99%forthe proposedalgorithm.Itisreducedby6.73%.

Table 2 Result Comparison of LinearDiscriminantAnalysiswithpropo sedalgorithm

Metrics	Linear Discriminan tAnalysis	Proposed algorithm
AccuracyScore	80.28%	87.01%
F1Score	88.83%	93.01%
PrecisionScore	97.67%	99.17%
Recall	81.45%	87.58%
MissRate	18.55%	12.42%
Fall-out	49.39%	57.36%
Specificity	50.61%	42.64%
MeanAbsoluteEr ror	19.72%	12.99%

Linear Discriminant Analysis is a classifiermodel tested with the proposed algorithm. Theresults,shownintheabovetable,werecompar edusingaccuracyscore,F1score,precisionscore,r ecall,missrate,fall-out,specificityandmeanabsoluteerror.Theresult sdemonstratethattheproposedalgorithmisimpro vingtheanomalydetectionperformance.

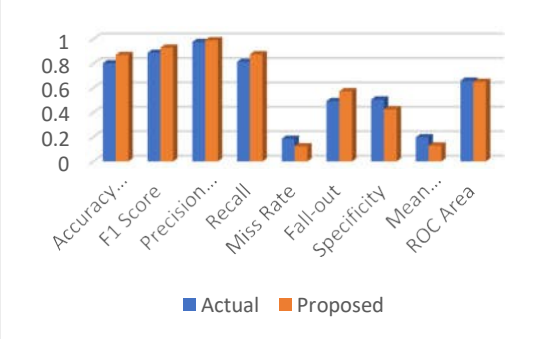

Figure 12 Performance Metrics of LinearDiscriminantAnalysiswithproposedalgor ithm

VI CONCLUSION

TheproposedalgorithmtestedwithQuadratic DiscriminantAnalysisandLinearDiscriminantA nalysismodels.ItimprovesAccuracyscore,Precis ionscoreandF1score.

QuadraticDiscriminantAnalysiswithpropose d algorithm. The Accuracy score hasimproved by 5.24%. The Precision score hasimprovedby12.53%.TheF1scorehasimprove dby7.91%.

Linear Discriminant Analysis with proposedalgorithm.TheAccuracyscorehasimpro vedby8.37%. The Precision score has improved by1.54%.TheF1scorehasimprovedby4.71%.

VII FUTUREWORK

The proposed algorithm can be used withdifferent classifiers. They are Gaussian NaiveBayes,Logistic Regression, MLP Classifier,Adaboost Classifier, Random Forest Classifier,DecisionTreeClassifier,andK-NearestNeighbors.

Trainthemodelswithdifferentdomaindatasets. The wide variety of training datasetspredicts with better accuracy. The dataset canbe retrieved from different seasons and differentgeographicallocationpeopledata.

REFERENCES

[1] BayuAdhiTama,LewisNkenyereye,S. M.RiazulIslam,andKyung-SupKwak(2020).An Enhanced Anomaly Detection inWebTrafficUsingaStackofClassifierEnsem ble.IEEEAccess,DigitalObjectIdentifier10.11 09/ACCESS.2020.2969428,pages24120–24134.

[2] WajdiAlhakami, Abdullah Alharbi, SamiBourouis,RoobaeaAlroobaea,andNizar Bouguila(2019).NetworkAnomalyIntrusionD etection Using a Nonparametric BayesianApproachandFeatureSelection.IEEE Access,*DigitalObjectIdentifier10.1109/ACC ESS.2019.2912115,* *pages*52181–52190.

[3] HaowenXu, Wenxiao Chen, NengwenZhao, Zeyan Li, Jiahao Bu, Zhihan Li, YingLiu, Youjian Zhao, Dan Pei, Yang Feng, JieChen,ZhaogangWang,HonglinQiao(2018) .UnsupervisedAnomalyDetectionviaVariatio nal Auto-Encoder for Seasonal KPIsinWebApplications,CornellUniversity.

[4] Ya Su, YoujianZhao,ChenhaoNiu,RongLiu,WeiSun, DanPei(2019).RobustAnomalyDetectionfor MultivariateTimeSeries through Stochastic Recurrent NeuralNetwork.Proceedingsofthe25thACMS IGKDDInternationalConferenceonKnowledg eDiscovery&DataMining

[5] VenkatN.Gudivada,AmyApon,andJunhu aDing(2017).DataQualityConsiderationsforB igDataandMachineLearning: Going Beyond Data Cleaning andTransformations.InternationalJournalonA dvancesinSoftware10.1,pages1-20.

[6] AlamZaib, TarigBallal, ShahidKhattak,andTareqY.Al-Naffouri(2021),ADoubly

RegularizedLinearDiscriminantAnalysisClassifierWithAutomaticParameterSelection.IEEEAccess,DigitalObjectIdentifier10.1109/ACCESS.2021.3068611,pages51343–51354.

[7] HoussemSifaou, AblaKammoun,andMohamed-SlimAlouini(2020),High-Dimensional Quadratic Discriminant AnalysisUnderSpikedCovarianceModel.IEEE Access,DigitalObjectIdentifier10.1109/ACCESS.2020.3004812, pages117313–117323.

[8] Chen Zhuang, Hongbo Zhao, Chao Sun,and Wenquan Feng(2019), Detection andClassificationofGNSSSignalDistortionsBased on Quadratic Discriminant Analysis.IEEEAccess,DigitalObjectIdentifier 10.1109/ACCESS.2020.2965617, pages25221–25236.

[9] V. Sankar, Dr. G. Zayaraz(2020). JSONandXMLdatasecurityperformanceofEncryptedFileProcessinWebServices.Published in International Journal of CurrentScience(IJCSPUB),Volume10,Issue4.

[10] Ravi Chandra Jammalamadaka, SharadMehrotra (2006). Querying Encrypted XMLDocuments. Published International DatabaseEngineeringandApplicationsSymposium(IDEAS'06)

[11] RA.K.Saravanaguru,GeorgeAbraham,Krishnakumar Venkatasubramanian,KiransinhBorasia(2013).SecuringWebServicesUsingXMLSignatureandXMLEncryption. School of Computer Science andEngineering,VITUniversity,Vellore,India.

[12] Hoi Ting Poon and Ali Miri(2015).ComputationandSearchoverEncryptedXML Documents. Department of ComputerScienceRyersonUniversityToronto, Ontario, Canada,978-1-4673-7278-7/15

[13] GuYue-sheng, Ye Meng-tao, Gan Yong(2010).WebServicesSecurityBasedon XML Signature and XML Encryption.Journalofnetworks,Vol.5, No.9.

[14] Nithin N and Anupkumar M Bongale(2012). XBMRSA:A New XML EncryptionAlgorithm. Proceedings of Information andCommunicationTechnologies.WorldCongress,pp567-571,2012.

[15] NithinNandHarshitha.K.S(2014).Analysis of Symmetric algorithm for XMLdocument security. International Journal ofInnovations in Engineering and Technology(IJIET),Vol.3Issue4.

[16] AamerNadeem (2005). A PerformanceComparison of Data Encryption Algorithms.IEEE.

[17] RaviVarma1,Dr.G.VenkatRamiReddy (2014). Schema Based Parallel XMLParser:AFastXMLParserDesignedforLarge XML Files. International Journal ofComputerScience andMobile Computing.Vol.3Issue.8.

[18] Shikha Agarwal, BalmukumdJha, TisuKumar,ManishKumar,PrabhatRanjan(2019). Hybridof NaiveBayesandGaussian Naive Bayes for Classification: AMap Reduce Approach. International JournalofInnovativeTechnologyandExploringEngineering(IJITEE)ISSN:2278-3075,Volume-8,Issue-6S3.

[19] RasimM.Alguliyev,RamizM.Aliguliyev and FarganaJabbarAbdullayeva(2019).Hybridisationofclassifiersforanomaly detection in big data. InternationalJournalofBigDataIntelligence.10.1504/IJBDI.2019.10018528