INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND APPLICATIONS
VOLUME X, ISSUE II, FEB 2016, WWW.IJCEA.COM, ISSN 2321-3469

OUTLIER DETECTION EFFICIENCY FOR HIGH DIMENSIONAL DATA

C.Jayaramulu[1], P.Krishna Akhila[2],M.Geetha Mounika[3],SK.Jasmine[4], B. Divya[5]

[1]Associate Professor, Krishna Chaitanya Institute of Technology &Sciences , Markapur,A.P, India
[2,3,4,5] Scholar, Krishna Chaitanya Institute of Technology &Sciences , Markapur, A.P, India

ABSTRACT:

A difficult problem in machine learning is still how to properly and efficiently handle huge dimensionality of data. Real-world applications for identifying abnormal items from provided data are numerous. The high-dimensional issue and the size of the neighbourhood in outlier discovery have not yet garnered enough attention, despite the fact that many traditional outlier detection or ranking algorithms have been seen during the previous several years. While the latter necessitates suitable parameter values, making models extremely complex and more sensitive, the former may lead to the distance concentration problem where the distances of observations in high-dimensional space tend to be indiscernible. We offer a notion termed local projection score (LPS) to express the degree of divergence of an observation from its neighbours in order to partially overcome these issues, particularly the high dimensionality. The low-rank approximation method is used to get the LPS from the neighbourhood information. An observation with a high LPS is likely to be an anomaly with good odds. Based on this idea, we provide an effective outlier identification technique that is also resistant to the k-nearest neighbour parameter. The performance of the suggested technique is competitive and promising, as demonstrated by extensive assessment trials using five well-known outlier identification methods on twelve publicly available real-world data sets.

Key Words—Dimension reduction ,high-dimensional data, k nearest neighbors(kNN),low rank approximation, outlier detection

[1] Introduction

A growing amount of data is being accessible in practical applications because to the development of developing technologies. Numerous factors, like as broken hardware or malicious activity, might cause part of the enormous data to exhibit odd behaviours or patterns. These outliers, ambiguities, abnormalities, novelties, or deviants—otherwise known as outliers, errors, abnormalities, novelties, or deviants—are

extraordinary behaviours or inconsistent patterns that deviate from the data's expected course [1], [2]. According to various objectives, they frequently appear in reality as representations of sounds or fascinating facts, such as cyber-intrusion and terrorist actions [3]. Because it may reveal unexpected behaviours, intriguing patterns, and extraordinary occurrences from data, identifying outliers from data is of tremendous interest to the communities of machine learning and data mining.In fact, in the preprocessing phase of data analysis, locating or removing outliers becomes crucial [4]. Because noises might prevent the finding of crucial information, noise reduction can, for instance, increase model performance. Anomaly access detection, on the other hand, can help us distinguish between intrusion and network access by looking at access records in a firewall one at a time. A procedure called outlier identification, sometimes referred to as anomaly detection, is used to identify unexpected findings that differ significantly from the majority of the observations [5]. Outlier detection has drawn a lot of interest and been used in a wide range of disciplines because it can significantly improve decision analysis, including criminal and terrorism identification [6], defect debugging and diagnosis [7], and many others Network intrusion, fraud detection, medical and health monitoring, signal analysis, image processing, anomalous weather detection, estimation of aberrant crowd behaviour, and many more fields [1], [2], [8],[9],[10],[20], and [21]. The large range of real-world uses confirms that outlier identification is a hotly debated subject. There is a substantial corpus of work on outlier identification techniques. Technically, there are two primary steps in the process of finding outliers: outlier ranking [17] and determination. The former provides a ranking list of the observations, each of which has a score based on predetermined metrics. If a higher number indicates a greater degree of variation or anomaly [19], the observations with the highest scores appear at the top of the list.By using the ranking list, the latter detects outliers. According to this viewpoint, outlier ranking is essential to processing and decision-making for closed-loop systems that have to keep vital physiological parameters, such oxygen level, within a predetermined range of values. detection. The two terms most frequently used in the literature—and sometimes used synonymously—are outlier ranking and outlier detection. Statistics-based, distance-based, density-based, and clustering-based methods are a few categories into which the outlier detection algorithms can be roughly divided [1]. Due to the intuitive nature of their concepts and ease of application, the distance-based and density-based detection methods have drawn the most attention and have been the subject of extensive study.The local outlier factor (LOF) [12] and the rank-based detection algorithm (RBDA) [11] are typical instances of these two categories, respectively. For outlier detection, there are two key issues that require additional research. The high dimensionality of data is the first one. The so-called "dimensionality curse" and the "distance concentration" are two prevalent issues that the high dimensionality may bring up [13]. While the latter indicates that the distance or density metrics fail to capture the neighbourhood information because all distances between observations tend to become indiscernible as the dimensionality increases, the former refers to the fact that the size of observations grows exponentially with the number of dimensions, making the data sparse.Dimension reduction is a typical approach to the high-dimensional problem in machine learning. For instance, Kasun et al. [14] used an extreme learning machine to design an effective dimension reduction approach. It is still challenging to accurately and efficiently separate outliers from typical data in a high-dimensional space. The second problem is that, despite their widespread use, distance-based detection techniques only consider global data and their effectiveness is influenced by the size of the neighbourhood, whereas density-based techniques are sensitive to neighborhood-specific features [1], [5]. In this research, we try to solve the aforementioned issues by creating a unique but powerful learning approach for outlier detection. The idea that abnormal findings have greater variances and differ considerably from other observations within the same neighbourhood of data serves as the basis for the suggested strategy. A brand-new statistic known as local projection score (LPS) is presented to measure the level of deviation. It is mostly used to quantify how far off each observation is from its corresponding neighbours when those neighbours are projected into a cheap space as a result of dimension reduction. It should be noted that LPS may handle high-dimensional data without any particular dimensionality restrictions in addition to taking local information into consideration. As a result, we are able to provide a rule of thumb for classifying and identifying outliers, according to which an observation with a high LPS value is likely to be an outlier with a high likelihood. Specifically, our method starts to identify k nearest neighbors(kNNs) for each observation.

[2] LITERATURE SURVEY

In biomedical applications of computational biology, such as applications like gene and SNP selection from high-dimensional data, biomarker identification is a crucial area. Surprisingly, only lately has the robustness or stability of such selection processes been discussed in relation to sample variance. The resilience of biomarkers is a crucial problem, though, since it might have a significant impact on later biological validations. Additionally, a more complete collection of markers could provide an expert greater reason to be confident in the outcomes of a selection procedure. A broad framework for evaluating the sustainability of a biomarker selection method is our first contribution. Second, we performed a thorough investigation of the recently developed ensemble feature selection approach, which combines numerous feature selections to enhance the robustness of the ultimate set of selected features. We concentrate on selection techniques that are integrated into support vector machine estimates (SVMs). SVMs are strong classification models that have demonstrated cutting-edge performance on a variety of biological data-based diagnostic and prognosis applications. For challenges involving gene selection, their feature selection extensions also produced good results.We demonstrate how ensemble feature selection strategies may significantly improve classification performance while also enhancing the resilience of SVMs for biomarker identification. The suggested technique is tested on four microarray datasets, which demonstrate up to a 30% increase in the robustness of the chosen biomarkers and an improvement of about 15% in classification accuracy. The stability enhancement with ensemble methods is most pronounced for small signature sizes (a few tens of genes), which is most pertinent for the design of a diagnosis or prognosis model from a gene signature.

Several supervised learning methods perform better when particular cases are used. These consist of distributed networks, categorization rules, and algorithms that develop decision trees. However, algorithms that solely employ particular instances to resolve incremental learning problems have not been the subject of analysis. In this research, we present an approach and methodology for generating classification predictions using just particular examples, known as instance-based learning. Algorithms for instance-based learning do not save a collection of abstractions produced from particular examples. The closest neighbour technique, which needs a lot of storage, is extended by this method. We explain how storage needs may be drastically decreased with, at best, little learning rate and classification [18] accuracy losses.The storage-reducing technique operates well on a number of real-world databases, but when the amount of attribute noise increases in training examples, its performance rapidly deteriorates. In order to differentiate noisy cases, we expanded it with a significance test. The performance of this expanded approach diminishes smoothly as the noise level rises and compares favorably to a noise-tolerant decision tree technique.

By simultaneously observing the expression levels of hundreds of genes, oligonucleotide arrays can give a comprehensive picture of the status of the cell. The development of methods for extracting usable information from the generated data sets is of interest. Here, we describe the use of a two-way clustering technique to examine a set of data that contains the expression patterns of several cell types. Affymetrix oligonucleotide arrays complementary to more than 6,500 human genes were used to assess the gene expression in 40 tumour and 22 normal colon tissue samples. Both the genes and the tissues were subjected to an effective two-way clustering technique, which revealed wide, cogent patterns that point to a high level of structure behind gene expression in these tissues. Core gulated gene families grouped together, as seen in the case of the ribosomal proteins. Even when the expression of individual genes differed only marginally

between the tissues, clustering also distinguished between malignant and non-cancerous tissue and cell lines and in vivo tissues. Thus, two-way clustering may be useful for categorizing tissues according to gene expression as well as for classifying genes into functional groupings.

The efficacy of the defibrillation therapy in automated devices depends on the early diagnosis of ventricular fibrillation (VF). Numerous detectors based on temporal, spectral, and time-frequency properties taken from the surface electrocardiogram (ECG) have been proposed, however they consistently perform poorly. To increase the effectiveness of detection, machine learning techniques have been employed to combine ECG characteristics across many domains (time, frequency, and time-frequency). But the potential use of several parameters in machine learning schemes has increased the demand for effective feature selection (FS) methods. In this paper, we present a novel FS approach built on bootstrap resampling (BR) and support vector machine (SVM) classifiers.We provide a backward FS approach that depends on assessing how the performance of the SVM changes when features are eliminated from the input space. According to a nonparametric statistic based on BR, this examination was completed. After conducting simulation tests, we use the AHA and MIT-BIH ECG databases to benchmark the performance of our FS method. Our findings demonstrate that, in simulated situations, the proposed FS algorithm outperforms the recursive feature removal approach and that, with a smaller feature set, the VF detector works better.
We demonstrate that, in general, the Fisher linear discriminant rule performs much worse than the 'naive Bayes' classifier, which assumes independent covariates, when the number of variables increases more quickly than the number of data.

[3] IMPLEMENTATION

1.1 MODULES:

1.1.1 **Admin**: Using a legitimate account and password, the admin must log in to this module. After successfully logging in, he may perform a number of actions, such as see all users, their information, and approve them, add documents and their data, add images and their details, and more. View all papers and photos that have been uploaded, along with their rankings View users' search history for documents and photos with time delay and accuracy, as well as time delay results charts for documents and images with relevant keywords, rankings of the documents and images, and documents and image accuracy in charts.

1.1.2 User: There are n numbers of users present in this module. Before performing some tasks, the user must first register. After successfully registering, the user may log in using a valid user name and password. Once logged in, the user can perform a number of actions, including viewing profile information, searching for documents and images based on content keywords, and viewing the results with search time delay and data classification ((documents and images in separate folders with their corresponding size order from large to small) and showing accuracy (no. of query retrieved/total doc or images)*100). display the whole time-delayed search history.

1.2 Screen shots

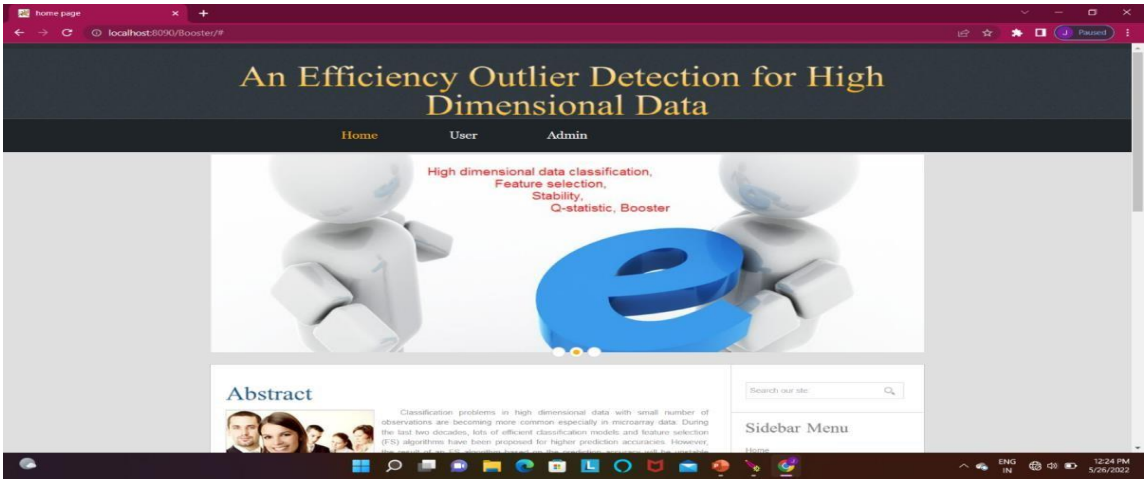

Fig. 1 Home Page

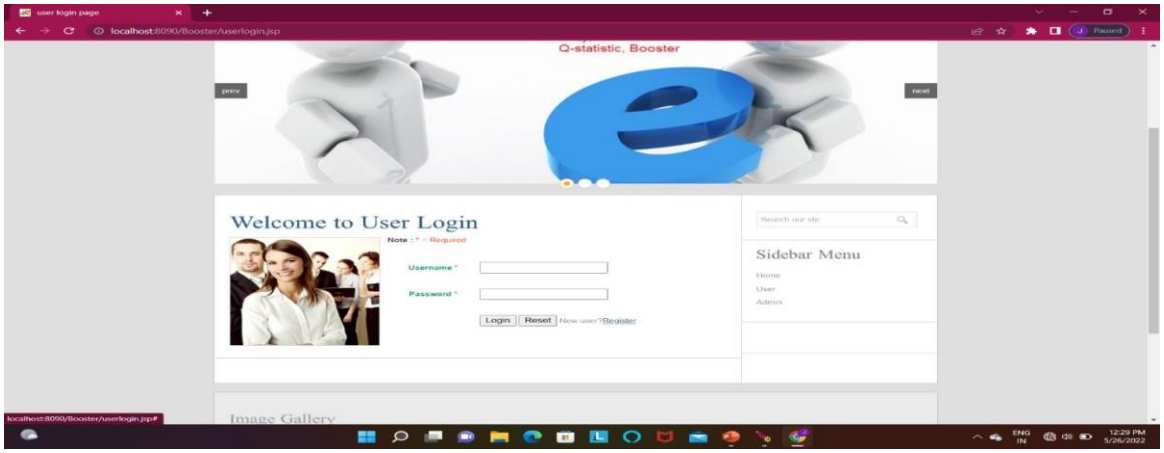

Fig. 2 User Login

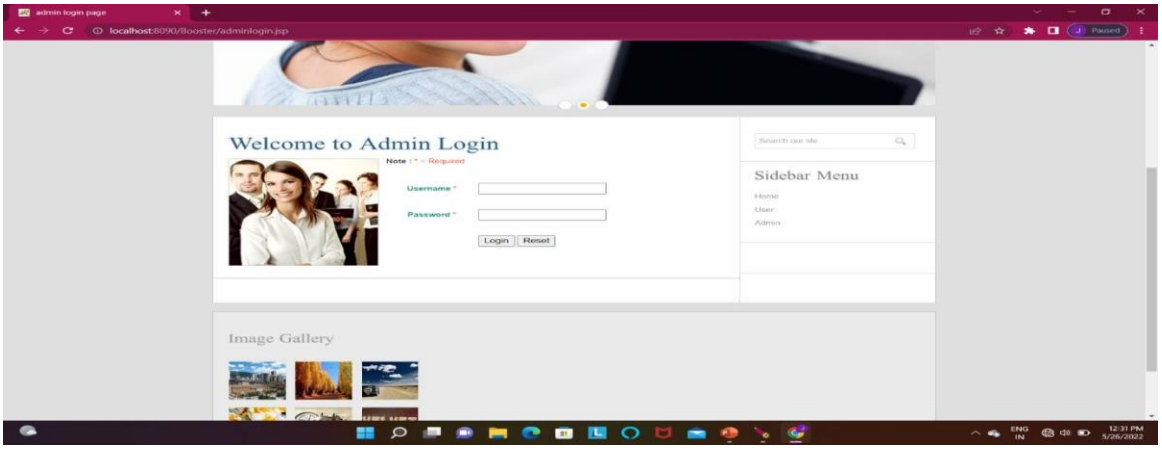

Fig. 3 ADMIN LOGIN

Above Screen Admin Login Page Will be displayed, It Contains User name and Password

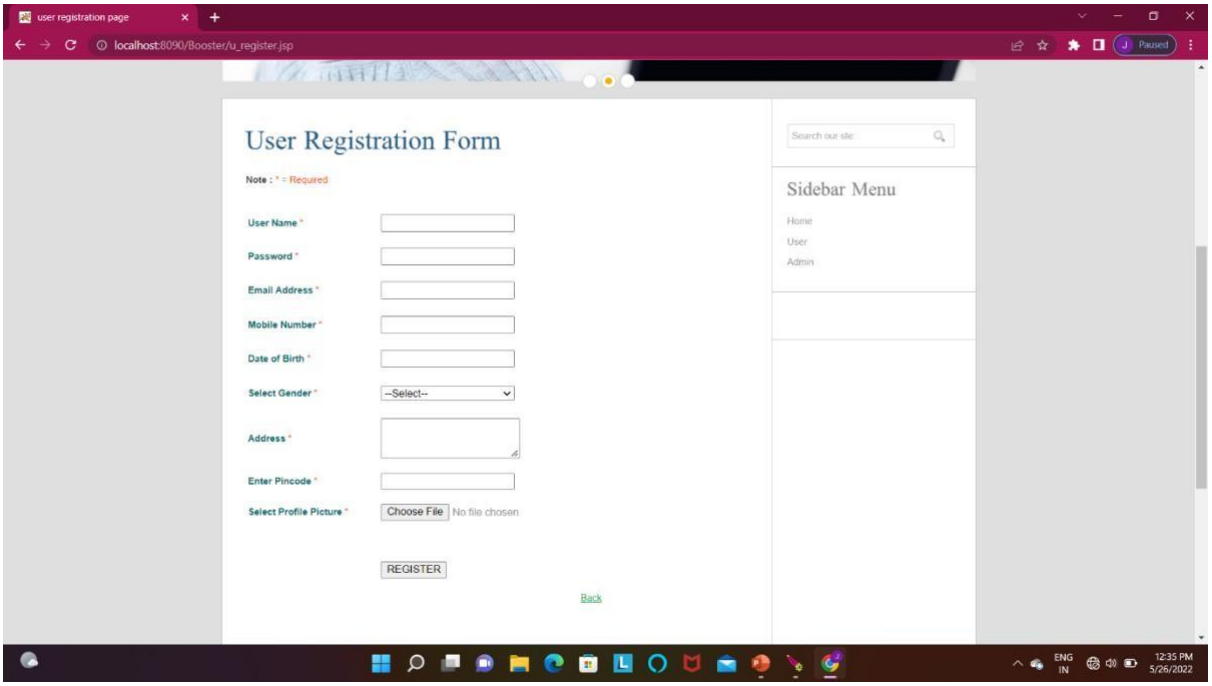

Fig. 4 Registration From

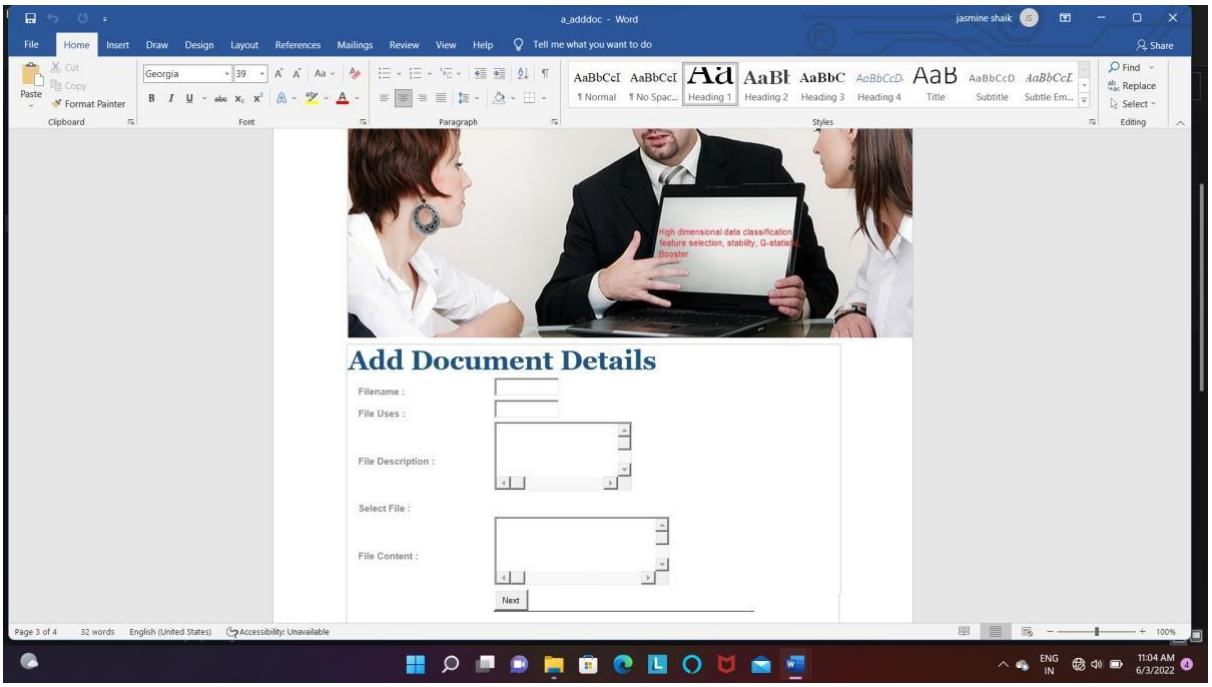

Fig. 5 Adding Document

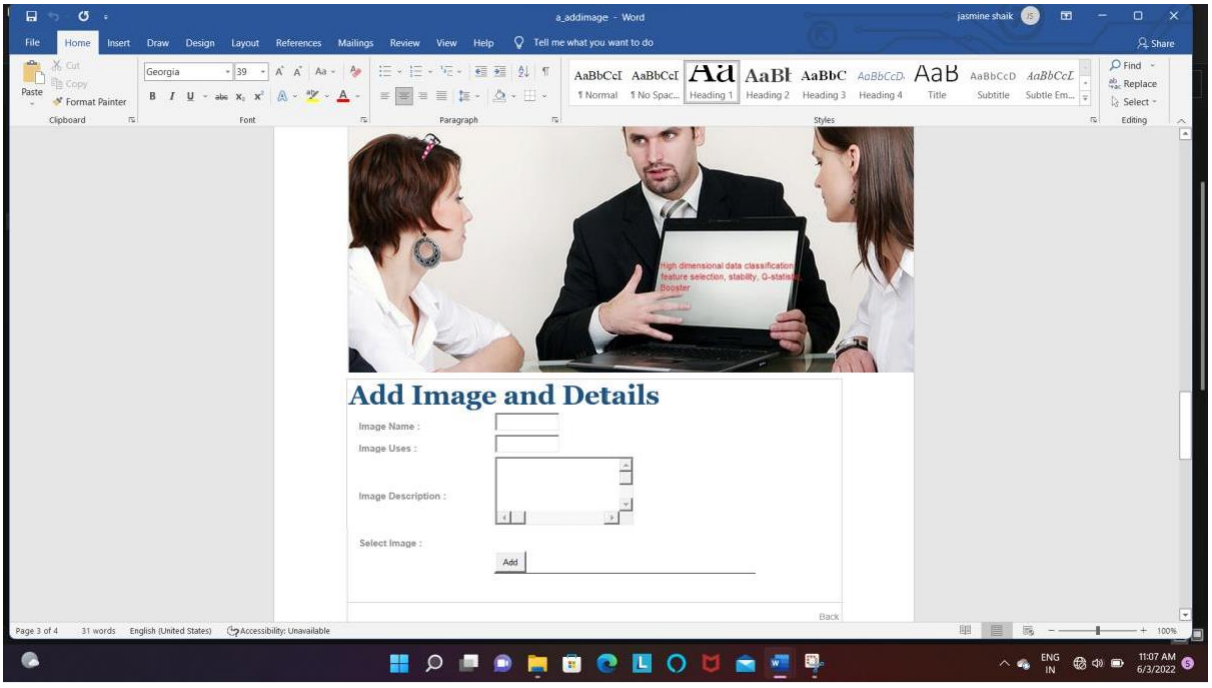

Fig. 6 Adding Image

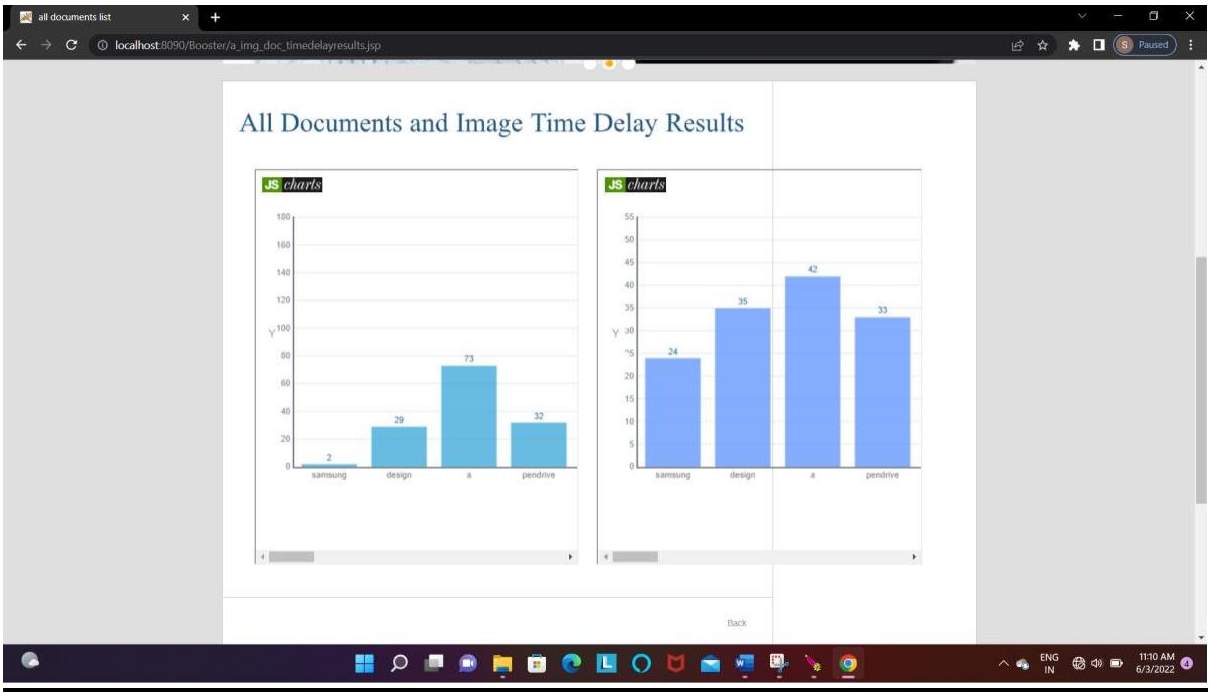

Fig. 7 All Documents And Image Time Delay Results

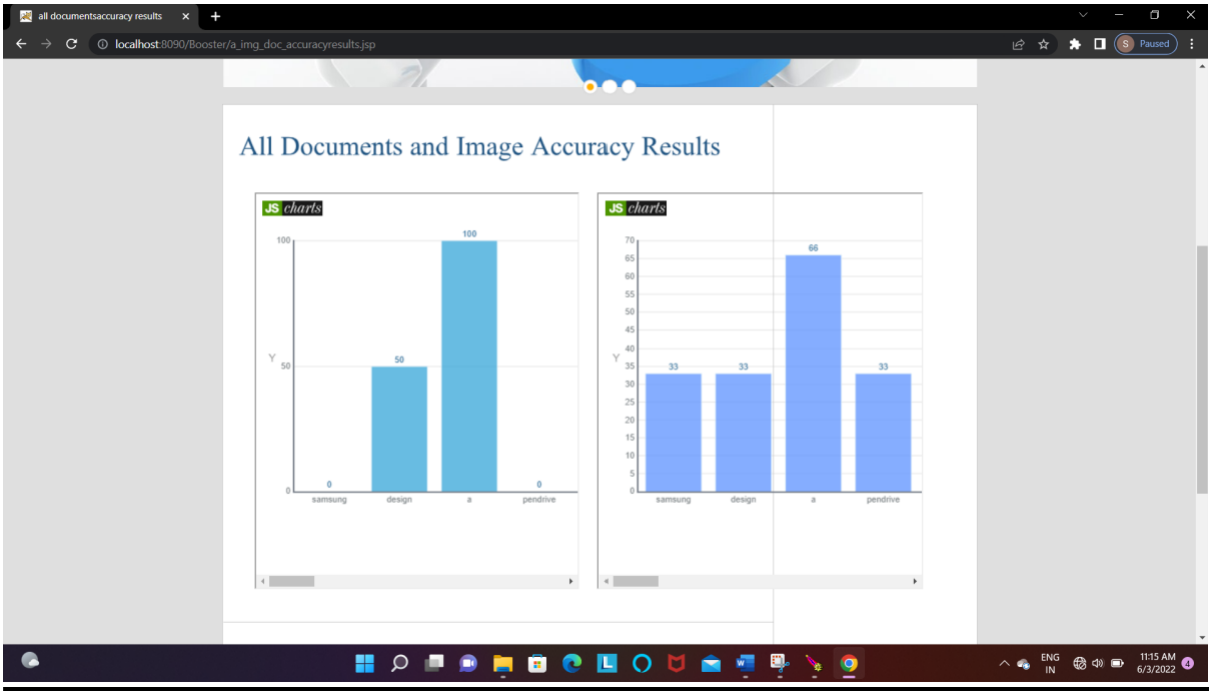

Fig. 8 All Documents and Image Rank Results

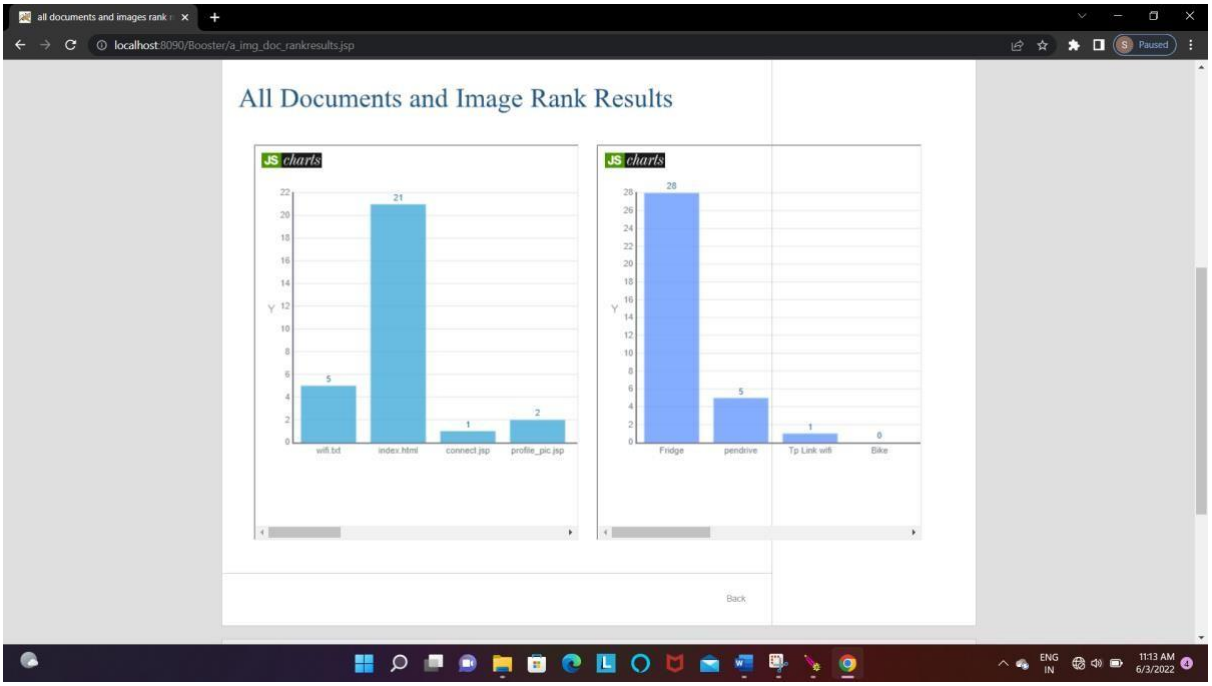

Fig. 9 All Documents and Image Accuracy Results

[5] CONCLUSION

We create an effective and quick learning system to recognise observations that deviate from the norm. The fundamental concept behind the suggested strategy is to use local neighbourhood information to assess an observation's outlier status. A concept known as LPS is presented to assess the abnormal degree of a suspicious observation in order to precisely capture the neighbourhood information. An observation

with a high LPS is likely to be an anomaly with good odds. Formally, the low-rank matrix approximation method can yield the LPS, which is compatible with the idea of nuclear norm. Furthermore, the suggested technique is resistant to the parameter k of the kNN contained within LPOD, in contrast to current distance-based and density-based detection methods.We conducted a thorough experiment using five well-known outlier identification methods on a variety of open real-world data sets to show the efficacy of our suggested approach. According to the experimental findings of the numerical comparison, the LPS is effective at identifying the most likely candidates to be outliers, and the performance of the LPOD is promising in many ways. Since LPOD uses kNN to obtain neighbourhood information, kNN is essential to its effectiveness, therefore the distance formulation of kNN will somewhat influence its performance. These factors will be taken into account in our future work when we extend LPOD to huge data scenario situations. Future research In this paper, we focus further on outlier identification methods for complex data with high dimensionality. The remainder of this essay is structured as follows. discusses cutting-edge outlier detection techniques for high-dimensional data, including the neighbour ranking-based approach. It offers the assessment metrics and datasets frequently used in outlier identification, then compares typical outlier detection techniques experimentally before discussing the difficulties and limitations of outlier detection in further research. The paper is finished at this point.

REFERENCES

[1]V.Chandola,A.Banerjee,andV.Kumar,"Anomalydetection:A survey,"ACMComput.Survey,vol. 41, no. 3, pp. 1–58, 2009.

[2]A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection inhigh-dimensionalnumericaldata,"Stat.Anal. DataMin.,vol. 5, no.5, pp. 363–387,2012.

[3]A. Kofuku and M. Georgopoulos, "A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes," Data Min. Knowl. Disc., vol. 20, no. 2, pp. 259–289, 2010.

[4] J. Ha, S. Seok, and J.-S. Lee, "Robust outlier detection using the instabilityfactor,"Knowl. Based Syst., vol. 63, no. 6, pp. 15–23, 2014.

[5]C.C.Aggarwal,OutlierAnalysis.New York, NY,USA:Springer,2013.

[6]V.RiffoandD.Mery,"Automateddetectionofthreatobjectsusingadaptedimplicitshapemodel,"IEEETrans.Syst., Man, Cybern.,Syst., vol.46,no. 4, pp.472–482, Apr. 2016.

[7]L.V.AllenandD.M.Tilbury,"Anomalydetectionusingmodelgenerationforevent-basedsystems without a preexisting formal model," IEEE Trans. Syst., Man, Cybern. A, Syst.,Humans,vol. 42, no. 3, pp. 654–668, May 2012.

[8]M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: Asurvey,"IEEETrans.Knowl. DataEng.,vol. 26, no.9, pp. 2250–2267,Sep. 2014.

[9]D.Berrar,"Learningfromautomaticallylabeleddata:Casestudyonclickfraudprediction,"Knowl. Inf. Syst., vol. 46, no. 2, pp. 477–490, 2016.

[10] R. Mitchell and I.-R. Chen, "Adaptive intrusion detection of malicious unmanned airvehicles using behavior rule specifications," IEEE Trans. Syst., Man, Cybern., Syst., vol. 44,no. 5,pp. 593–604, May 2014.

[11] H.Huang,K.Mehrotra,andC.K.Mohan,"Rank-basedoutlierdetection,"J.Stat.Comput.Simulat.,vol. 83, no. 3, pp. 518–531, 2013.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers, "inProc. ACMSIGMOD Int. Conf. Manag.Data, 2000,pp. 93–104.

[13] M. Radovanovic, A. Nanopoulos, and M. Ivanovi ´ c, "Reverse nearest ´ neighbors inunsupervised distance-based outlier detection," IEEE Trans. Knowl. Data Eng., vol. 27, no. 5,pp. 1369–1382,May 2015

[14] L.L.C.Kasun,Y.Yang,G.B.Huang,andZ.Zhang,"Dimensionreductionwithextremelearningmachine,"IEEETrans. ImageProcess.,vol. 25,no. 8, pp.3906–3918, Aug.2016.

[15] E.Eskin,"Anomalydetectionovernoisydatausinglearnedprobabilitydistributions,"inProc.17th Int. Conf.Mach. Learn., 2000, pp. 255–262.

[16] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical

outlierdetectionusingdirectdensityratioestimation,"Known.Inf.Syst.,vol.26, no.2,pp.309–336,2011.

[17] SkAlthaf Hussain Basha, VenkataPavan Kumar Savala, Ranganath P, P V Ravi Kumar, "Information Inclusion: The Modern Rank and The Approach Forward", International Journal of Computer Engineering and Applications(IJCEA), Volume 13, Issue 6 , December. 20, ISSN2321-3469.

[18] SK Althaf Hussain Basha, Ayesha Mariyam, and S Vishwanadha Raju "Applications of Multi- Label Classification", International Journal of Innovative Technology and Exploring Engineering(IJITEE),pp.86-89,ISSN:2278-3075, Volume-9 Issue-4S2, March 2020,RetrievalNumber:D10080394S220/2020©BEIESP,DOI:1035940/ijitee.D1008.0394S220,Blue Eyes Intelligence Engineering & Sciences Publication.

[19] Sk. Althaf Hussain Basha ,Mogili BVK Chaitanya Kumar, S VenkataPavan Kumar, "Distributed Anomaly Feature Detection Over Financial Frauds", International Journal For Recent Development In Science And Technology(IJRDST), Volume 4, Issue 1, Jan 2020, pp. 135-141, ISSN 2581 –4575.

[20] SkAlthaf Hussain Basha, Ch. Prakash, D. Mounika, G. Maheetha, "An Approach for Multi Instance Clustering of Student Academic Performance in Education Domain", IIJDWM Journal, Volume 3,Issue 1,pp.1-9,Feb.2013,ISSN:2249-7161

[21] SK Althaf Hussain Basha, NagaRaju Devarakonda, Shaik Subhani," Outliers Detection in Regression Analysis using Partial Least Square Approach", ICT and Critical Infrastructure: proceedings of the 48th Annual Convention of Computer Society of India- Springer, Vol II Advances in Intelligent Systems and Computing, Volume 249, pp. 125-135, Visakhapatnam, December 2013,ISBN: 978-3-319-03095-1.