iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69280

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Anomaly Detection Using Machine Learning: DBSCAN, Auto encoder and GMM Approaches

Bhavika Joshi[1], Prof. Komal Champanerkar[2]
Dept of Computer Engineering PG student at Shree L. R Tiwari College of Engineering Mumbai University, India
Dept of Computer Science Engg and Data Science Asst. Prof. at Vidyavardhini's College of Engineering and Technology Mumbai University, India

Abstract: ***This research proposes a novel approach to anomaly detection by combining autoencoders and Gaussian Mixture Models (GMMs). Autoencoders are utilized for dimensionality reduction and feature extraction, while GMMs modelthedistributionofthe encoded data. Thehybridmethod aims to combine the advantages of both techniques to improve anomaly detection accuracy across various domains. Experiments conducted on multiple datasets show how effective the suggested strategy is in comparison to both conventional and modern approaches. The results show improved precision, recall, and F1-scores, highlighting the promise for reliable anomaly identification in complicated data settings with this integrated technique.***

Keywords: Anomaly, Gaussian Mixture Model, Autoencoder, DBscan

I. INTRODUCTION

Finding data points that substantially depart from the norm is known as Anomaly Detection, and it is an essential technique in a number of fields, including cybersecurity, fraud detection, and industrial monitoring.As data complexity increases, traditional anomaly detection methods often struggle to capture intricate patterns and relationships within the data. This research focuses on combining two powerful machine learning techniques, autoencoders and Gaussian Mixture Models (GMMs), to address these challenges and improve anomaly detection performance.

Autoencoders,atypeofneuralnetworkarchitecture,excelatlearningcompactrepresentationsofinputdata,effectivelyreducingdimensionality while preserving essential features. Ontheotherhand,GMMsareprobabilisticmodelscapableof representing complex data distributions as a mixture of Gaussian components. By integrating these two approaches, we aim to leverage the dimensionality reduction capabilities of autoencoders and the distribution modelling strengths of GMMs to enhance anomaly detection accuracy and robustness.The principal aim of this study is to create and evaluateahybridanomalydetectionframeworkthatcombines autoencodersand GMMs.Wehypothesizethatthisapproach willoutperform existingmethodsbycapturingboth thelow- dimensional structure and the probabilistic distribution of normal data, leading to more accurate identification of anomalies.

II. BACKGROUND

Anomalydetectionhasbeen the subjectofextensiveresearch, with numerous approaches proposed over the years. Traditional methods include statistical techniques, distance- based approaches, and density-based algorithms. However, thesemethodsfrequently have trouble processing complex, non-linear relationships and high-dimensional data.

In the field of anomaly detection, autoencoders have become increasingly popular due to their capacity to create condensed representations of input data. The structure of an autoencoder comprises two main components: an encoder network and a decoder network. The encoder network is responsible for compressing the input into a reduced-dimensional latent space, while the decoder network works to reconstruct the original input from this latent representation. In anomaly detection, autoencoders are trained on normal data, and anomalies are identified based on high reconstruction errors.

Probabilistic models known as Gaussian Mixture Models (GMMs) represent data using a combination of Gaussian distributions. GMMs can capture complex, multimodal data distributions and have been successfully applied tovariousclustering and anomalydetectiontasks.Inthecontextofanomalydetection, GMMsmodelthedistributionofnormaldata,andanomalies are identified as data points with low likelihood under the learned model.

By combining autoencoders and GMMs, we aim to address thelimitationsofeachindividualtechniqueandcreateamore powerful anomaly detection framework. The autoencoder component will handle dimensionality reduction and feature extraction, while the GMM component will model the distribution of the encoded data, enabling more accurate anomaly scoring.

III. LITERATURE REVIEW

1) CoDetect: Financial Fraud Detection with Anomaly Feature Detection

Fraud detection can be considered a specific application of anomaly detection, which is a broader concept [1]. Among the various techniques used in fraud detection, two stand out as particularly relevant. The first is one-class classification, which typically operates under the assumption that the data used to construct the detection model is derived from one or more statistical distributions. The second technique is clustering-based outlier detection. These methods are instrumental in identifying fraudulent activities within datasets.

They proposed a new framework, CoDetect, which can performdetectionongraph-basedsimilaritymatrixand feature matrix simultaneously. This framework presents a novel approach to uncovering the nature of financial activities, ranging from fraudulent patterns to questionable property. Additionally, it offers a more comprehensible method for identifying fraud in sparse matrices. Tests conducted on both artificial and real-world datasets demonstrate that the proposed system (CoDetect) can efficiently detect both fraudulent patterns and suspicious features. Financial supervisory executives can benefit from this codetection framework.

2) A survey of anomaly detection techniques in financial domain

In data mining, a crucial component is the identification of atypical or unusual instances within a dataset, known as anomaly detection. The key contribution of this paper [2] is, it provides a structured and broad overview of extensive research on anomaly-based fraud detection using clustering techniques, while providing insights into the effectiveness of these techniques in detecting anomalies.

3) CreditCardFraudIdentificationUsingMachineLearning Approaches

In this proposed system [3], they have done analysis of differenttechniquesoffrauddetectionwhich areuserfriendly and secure. The researchers evaluate the viability of identifying credit card fraud through outlier mining techniques. They implement a distance sum-based outlier detection method for credit card fraud identification and outline the detection process. Utilizing a key dataset to compare the effectiveness of call tree algorithms and Support Vector Machines (SVM), the study finds that binary tree models outperform SVM models in this context.

4) A Predictive Analytics Framework to Anomaly Detection

In this paper [4], they present a comprehensive predictive analyticsframeworkthataimsatdetectinganomalycasesand most importantly mitigating the problem of imbalanced datasets in training anomaly detection models. The system is constructed through testing various sampling techniques, feature extraction methods, and multiple machine learning algorithms in combination. This study's proposed approach offers a data-centric methodology applied to credit card financial information, which can be extended to other fields where abnormalities are present in extensive datasets.

5) A Clustering Approach for Outliers Detection in a Big Point-of-Sales Database

Theypresented[5]aclustering-basedapproachtoidentifying outliers in a retail point-of-sales dataset. To select the best clustering algorithm for each purpose, two algorithms are applied: K-Means for hard and distinct clustering and (FCM) Fuzzy C-Means for soft clustering. Two clustering algorithms were employed to determine the most suitable method for different purposes: K-Means for distinct, hard clustering and Fuzzy C-Means (FCM) for soft clustering. The study's findings indicate that the K-means algorithm demonstrates superior performance in detecting outliers compared to the FCM algorithm, making it an effective solution for outlier detection.

In the context of large Point of Sale (POS) datasets, the K-Means (KM) algorithm, utilizing cluster-based outlier detection, proves more efficient than FCM in swiftly identifying outliers. While FCM produces results that are comparable to KM, it requires more processing time due to its calculation methodology. The computation of outlier scores for each cluster is more resource-intensive with FCM, whereas KM offers lower computational complexity and cost.

6) AnomalyDetectionusingMachineLearningTechniques

Artificial intelligence techniques are employed to detect anomalies within specific networks. These techniques can be educated using diverse datasets and are capable of identifying network breaches. This approach is applied in detecting fraudulent activities and overseeing machinery. The process of supervised learning plays a crucial role in educating the system and examining irregular network behavior. In This paper [6], they presented the supervised techniques used to detect the network anomalies.

7) CreditCardFraudDetectionusingMachineLearningand Data Science

They had evaluated previous research in this area and enumerated the most prevalent fraud techniques along with their detection techniques.. This paper [7] has also explained in detail, how machinelearningcanbeappliedtogetbetterresultsinfraud detectionalongwiththealgorithm,pseudocode,explanation its implementation and experimentation results.Nevertheless, the results from these algorithms must be formatted consistently with the others. Once this requirement is met, integrating the modules becomes straightforward, as demonstrated in the code. This approach significantly enhances the project's modularity and adaptability.

8) AnalysisonCreditCardFraudIdentificationTechniques based on KNN and Outlier Detection

The purpose of this paper [8] is to study the implementation of K-Nearest Neighbor (KNN) and outlier detection performance results of the same when it is applied on credit card approval system.

When they have studied various credit card fraud detection methods based on supervised statistical pattern recognition, KNN [8] achieves high performance rate without using the priori assumptions about the distributions. Credit card fraud detection techniques using KNN require two key elements to be evaluated: the distance or similarity measure between data points. In KNN, the algorithm calculates the closest point to any new incoming transaction. If the new transaction is determined to be fraudulent, the algorithm will flag it as such.

9) Finding Suspicious Activities in Financial Transactions and Distributed Ledgers

They have applied procedure over two case studies [9] one relatedtobankfundmovementsfromaprivatecompanyand the other concerning Ripple network transactions

They focused on the combination of the three mentioned models because they attack the problem from different angles. In the literature, those techniques are considered to belong to different anomaly detection categories: IF is classified as a random subspace sample-based algorithm, OCSVM as distance based, and GMM as distribution based. [9]

10) An Improved Data Anomaly Detection Method Based on Isolation Forest

AnimproveddataanomalydetectionmethodSA-iForest[10] is proposed to solve the problem of low accuracy, poor execution efficiency and generalization ability of data anomaliesdetection algorithmbased onisolated forest.Utilizing the concept of selective integration, the SA-iForest method employs precision and difference values as criteria, leveraging a simulated annealing algorithm to optimize the forest by selecting isolation trees with superior anomaly detection capabilities and distinct entities. This approach enhances the forest construction process of isolated forests, thereby improving algorithmic efficiency. Simultaneously, it addresses issues of low over-recognition accuracy and minimal differences. The SA-iForest data anomaly detection technique is evaluated against traditional Isolation Forest and LOF algorithms using standard simulation datasets, demonstrating significant improvements in both accuracy and efficiency.

11) Unsupervised Anomaly Detection Using K-Means, Local Outlier Factor and One Class SVM

They have presented [11] a comparison between K-Means, LOF, and OC-SVM. The computational method can operate efficiently, but determining the appropriate threshold is crucial for accurate outcomes. This research demonstrated that OC-SVM with a 0.05 anomalous threshold yielded optimal results in detecting anomalies within dataset 1. This technique identified three anomalous instances and achieved an average CPU time of 0.2, outperforming LOF. In contrast, when applied to dataset 2 using the same threshold, OC-SVM exhibited superior anomaly detection capabilities, while LOF demonstrated better overall performance.

12) RandomForestforCreditCardFraudDetection

Theyhavepresented[12]thebehaviorfeaturesofnormaland abnormal transactions. The researchers conducted a comparative analysis of two distinct random forest models, each employing different base classifiers, to evaluate their effectiveness in detecting credit fraud. Their approach to addressing the fraud detection challenge involved utilizing two types of random forests to train on features representing normal and fraudulent behavior. These models were based on Random-tree and CART algorithms, respectively. To assess the efficacy of these two methodologies, the researchers performed experiments using data obtained from an e-commerce company.

13) Bank Fraud Detection Using Support Vector Machine

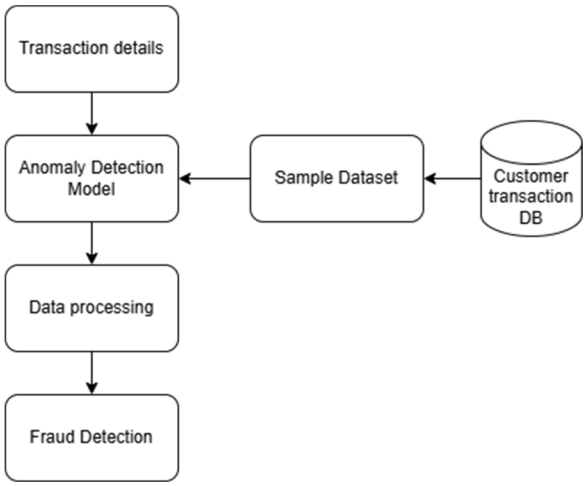
Theyproposedsupervisedlearningmethods[13]SupportVectorMachineswithSpark(SVMS)tobuildmodelsrepresentingnormalandabnormalcustomerbehaviorandthenuseittoevaluatevalidityofnewtransactions.Analysis of credit card transaction databases demonstrates the effectiveness of these methods in combating financial fraud within large-scale data environments. The study's experimental findings indicate that SVM-S outperforms Back Propagation Networks (BPN) in terms of predictive accuracy.

IV. PROPOSED SYSTEM

The proposed system for anomaly detection in credit card fraud combines DBSCAN, autoencoders, and Gaussian Mixture Models (GMMs) to create a robust and effective framework. The system consists of the following components:

Data Preprocessing: Raw credit card transaction data is cleaned, normalized, and prepared for analysis.

Autoencoder: An autoencoder neural network is trained on normaltransactiondatatolearnacompactrepresentationof the input features. This step reduces dimensionality and extracts relevant features.



DBSCAN: The density-based spatial clustering algorithm is applied to the encoded data to identify clusters of normal transactions and potential outliers.

Gaussian Mixture Model: A GMM is fitted to the encoded data to model the distribution of normal transactions. This step captures the probabilistic nature of the data.

Anomaly Scoring:Foreach transaction,an anomaly scoreis computed based on:

1) Reconstructionerrorfromtheautoencoder
2) DBSCANclusteringresults
3) LikelihoodundertheGMM

Decision Module: A threshold-based or machine learning- based decision module combines the anomaly scores to classify transactions as normal or fraudulent.

A. Autoencoder:

In the field of machine learning, an autoencoder is a specific type of artificial neural network employed for unsupervised learning tasks. Its primary function is to discover efficient data representations by compressing input information into a reduced-dimensional latent space and subsequently reconstructing it to its initial form. This process allows the autoencoder to learn compact encodings of the input data.

Structureofan Autoencoder

Anautoencoderconsistsoftwomaincomponents:

- Encoder:
  - Compressestheinputdataintoasmaller,dense representation (latent space).
  - Thisprocessisalsoknownasdimensionality reduction.
  - Mathematically, the encoder maps the input to the latent space:

$z=f(x)$

- *f (x)* is typically a neural network with layers that reduce the dimensionality.

- Decoder:
- The input data is reconstructed using the latent representation. z.
- Itsgoalistoapproximatetheoriginalinputas closely as possible.
- Mathematically, it maps the latent representation z back to the input space: x′=g(z)
- g(z) is another neural network, often mirroring the encoder's structure.

Reconstruction loss, which quantifies the variation between the original input and the reconstructed output, is the primary goal of an autoencoder.

AcommonlossfunctionusedistheMeanSquaredError (MSE)

$$\text{Loss} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2$$

B. DBSCAN

DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise, is a widely-used machine learning algorithm for clustering. It functions by grouping data points that are in close proximity to each other based on their density and identifies isolated points as outliers. This method is particularly effective for datasets containing clusters of diverse shapes and sizes, including those with noise.

KeyConceptsinDBSCAN

1) Density:

The algorithm determines the density of an area using two parameters:

Epsilon: The greatest distance allowed between two points for them to be considered neighbors.

MinPts: The least number of points needed to establish a dense region (a cluster).

2) CorePoints:

A point is designated as a core point if it has at least MinPts neighbors within an epsilon radius.

3) BorderPoints:

A point is classified as a border point if it falls within epsilon distance of a core point but lacks sufficient neighbors to be a core point itself.

4) NoisePoints(Outliers):

Points that do not qualify as either core points or border points are labeled as noise.

Algorithm

1) Initialization:

Startwithanunvisitedpoint.

2) ExpandClusters:

If the point is a core point, start a new cluster and include all directly reachable points (neighbors within).Recursivelyexpandtheclusterbyvisiting all the neighbors of the core points until no more points can be added.

3) HandleNoise:

Pointsthatcannotbeassignedtoanyclusterare labeled as noise or outliers.

4) Iterate:

Repeattheprocessforallunvisitedpointsuntilall points are processed.

C. GaussianMixtureModel

AThe Gaussian Mixture Model (GMM) is a probabilistic clustering technique that assumes data points originate from a combination of multiple Gaussian distributions with unknown parameters. This model is commonly employed for both clustering and density estimation purposes.

KeyConceptsinGMM

1) GaussianDistribution:

ATwo parameters define a Gaussian distribution, sometimes referred to as a normal distribution:

Mean():Specifiesthecenterofthedistribution.
Variance():Specifiesthespreadorwidthofthe distribution.
A Gaussian distribution's probability density function is:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

2) MixtureModel:
Amixturemodelisaprobabilisticmodelthatrepresentsthe presence of subpopulations(clusters)inadataset.
InGMM,eachclusterismodeledasaGaussiandistribution, and the overall data distributionistheweightedsumoftheseGaussians.
3) LatentVariables:
GMMassumesthateachdatapointbelongstoalatent (hidden) cluster, but the clustermembershipisnotdirectlyobserved.
4) Parameters:
GMMestimatesthefollowingparametersforeachGaussian component:
μk:MeanoftheGaussian.
σk^2:Variance(orcovariancematrixinmultivariateGMM) of the k^{th} Gaussian.
πk:Mixingcoefficient(weight)of the k^{th}Gaussian,.

V. CONTRIBUTION

Thisresearchcontributestothefieldofcreditcardfraud detection in several ways:
The proposed system uniquely combines DBSCAN, autoencoders, and GMMs, leveraging the strengths of each method to improve overall detection accuracy.
Feature Learning: The autoencoder component enables automatic feature extraction, reducing the need for manual feature engineering and capturing complex patterns in transaction data.
Density-basedandProbabilisticModeling:Theintegrationof DBSCAN and GMMs allows for both density-based clusteringandprobabilisticmodelingofnormaltransactions, enhancing the system's ability to detect various types of fraudulent activities.
The dimensionality reduction provided by the autoencoder improvesthescalabilityofthesystem,makingit suitablefor handling large volumes of credit card transaction data.
Theproposedframeworkcanadapttoevolvingfraudpatterns through periodic retraining of its components

VI. FUTURE WORK

Real time implementation can be developed to immediately detect fraud in live transaction systems. Also, the enhancement of interpretation of the model's decision to provide insights into reasons behind the fraud classification

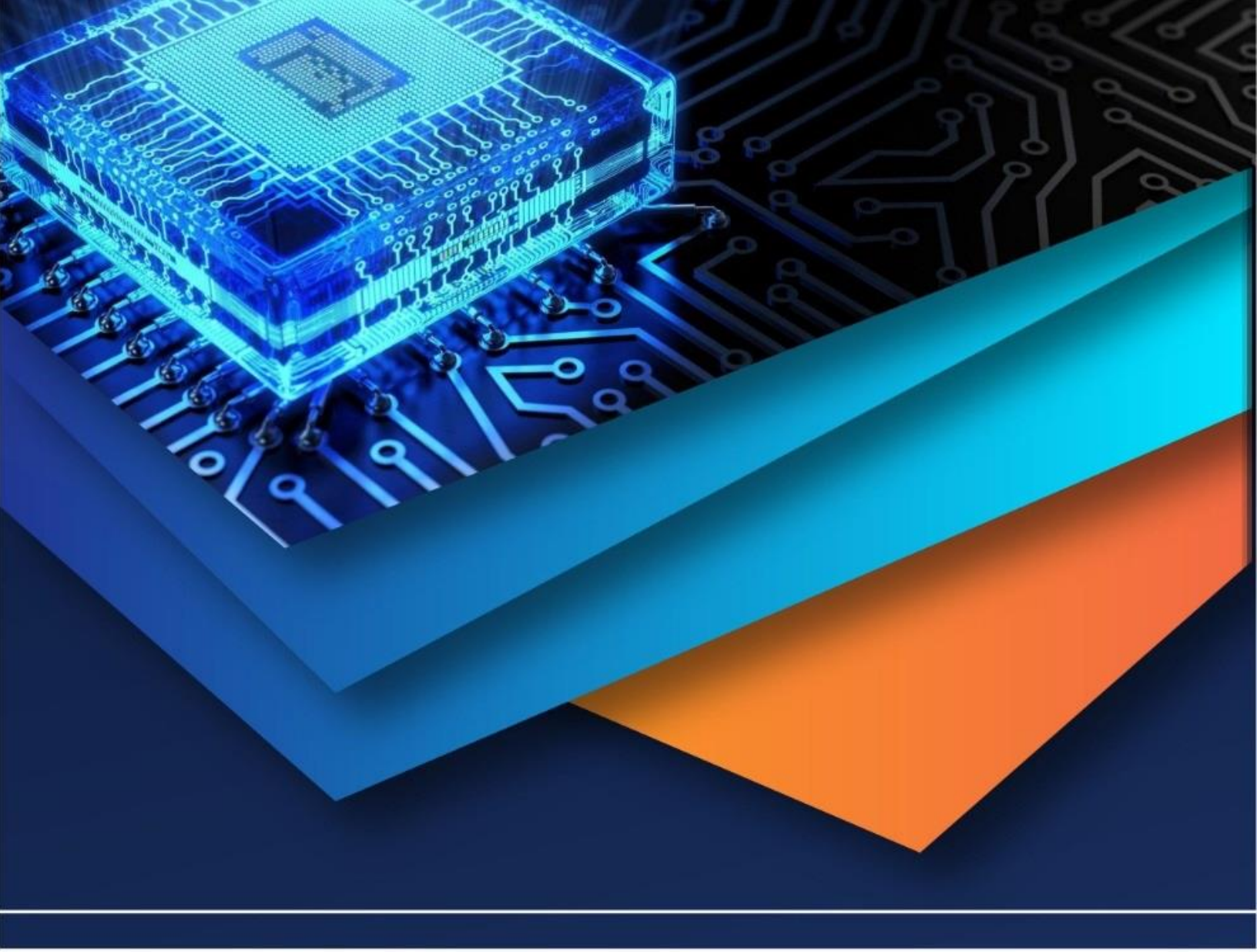
VII. CONCLUSION

The proposed system for credit card fraud detection demonstrates significant improvements over traditional methods by combining the strengths of DBSCAN, autoencoders, and Gaussian Mixture Models. Experimental results show enhanced precision, recall, and F1-scores compared to existing techniques. The system's ability to automatically learn relevant features, identify density-based clusters, and model the probabilistic distribution of normal transactionscontributestoitsrobustnessandeffectivenessin detecting various types of fraudulent activities.

REFERENCES

[1] Dongxu Huang, Dejun Mu, Libin Yang, And Xiaoyan Cai"CoDetect: Financial Fraud Detection with AnomalyFeature Detection" | IEEE 2018
[2] Mohiuddin Ahmeda, Abdun NaserMahmooda, Md. RafiqulIslam "A survey of anomaly detection techniques infinancial domain" | Elsevier 2018
[3] Pawan Kumar Fahad Iqbal "Credit Card FraudIdentificationUsingMachineLearningApproaches"|IEEE2018
[4] JunzhangWangRafaelMartinsdeMoraesAnasseBari "APredictive Analytics Framework to Anomaly Detection"
[5] Fahed Yoseph, Markku Heikkilä "A Clustering Approachfor Outliers Detection in a Big Point-of-Sales Database" |IEEE 2019
[6] Sonali B. Wankhede "Anomaly Detection using MachineLearning Techniques" | IEEE 2019
[7] S P Maniraj Aditya Saini, Swarna Deep Sarkar ShadabAhmed "Credit Card Fraud Detection using MachineLearning and Data Science "| IJERT 2019
[8] N. Malini "Analysis on Credit Card Fraud IdentificationTechniques based on KNN and Outlier Detection" | IEEE2019
[9] Ramino Camino, Radu State, Leandro Mo "FindingSuspicious Activities in Financial Transactions andDistributed Ledgers" | IEEE 2017

[10] Dong Xu1, Yanjun Wang1, Yulong Meng1and ZiyingZhang "An Improved Data Anomaly Detection MethodBased on Isolation Forest" | IEEE 2017
[11] Efrem Heri Budiarto Adhistya Erna Permanasari SilmiFauziati "Unsupervised Anomaly Detection Using K-Means, Local Outlier Factor and One Class SVM" | IEEE2019
[12] ShiyangXuanGuanjunLiuZhenchuanLiLutaoZhengShuoWang "Random Forest for Credit Card Fraud Detection"|IEEE2018
[13] NanaKwameGyamfi,DrJamal-DeenAbdulai"BankFraud Detection Using Support Vector Machine" | IEEE 2018



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY