Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Real-Time Unsupervised Anomaly Detection in High-Dimensional Financial Data Streams

Sanika Thete
Department of Data Science
Dr. D. Y. Patil Arts, Commerce & Science College,
Pimpri,, Pune, Maharashtra, India

Abstract - The way digital financial systems are growing fast is creating a lot of transaction data all the time. This makes it really hard to find activity as it happens. Finding transactions is tough because there are so many more normal transactions than fake ones the way people make transactions is always changing and we do not have a lot of examples of fake transactions to learn from. This paper is about a system that can find activity in financial data as it happens without any help from people. It is made for data that has a lot of details. The system is good at finding activity in real-time and it does not need any help from people to do its job. The digital financial systems are growing and so is the transaction data that is why we need a system, like this to find transactions. The method they are suggesting uses something called Incremental Principal Component Analysis to figure out what normal transactions look like and find the ones that're not normal. It does this by looking at how it can rebuild the transactions. This way it can handle a lot of transactions coming in one after the other without using much computer power. They also used something called an Isolation Forest model to compare the results. They tried this method with a set of credit card transactions that they made look like they were coming in one at a time. They looked at how it worked by checking things like how often it was right how often it found the real anomalies and how well it could tell the difference between normal and not normal transactions. They used metrics, like accuracy and precision and recall and F1-score and ROC-AUC to evaluate the performance of the proposed method. The results demonstrate that the proposed approach achieves high precision and competitive detection performance while maintaining low latency, making it suitable for real-time financial anomaly detection applications.

Keywords

Unsupervised Anomaly Detection, Financial Data Streams, High-Dimensional Data, Incremental PCA, Credit Card Fraud Detection, Real-Time Analytics

INTRODUCTION

The rapid expansion of digital banking, online payments, and electronic commerce has led to the generation of massive volumes of financial transaction data in real time. Financial institutions are required to continuously monitor these transactions to detect anomalous or fraudulent activities that may result in significant financial losses. Real-time anomaly detection in financial systems is therefore a critical task; however, it presents several challenges due to the high dimensionality of transaction features, dynamic customer behavior, and strict latency requirements.

Financial anomaly detection is really tough because there are many more normal transactions than fraudulent ones, in real life. Fraudulent transactions are a tiny part of all the transactions that happen.

The old way of teaching computers to find these anomalies does not work well. This is because it needs lots of examples from the past where transactions were labeled as normal or fraudulent. It also assumes that people will keep cheating in the ways they did before.The problem is that we do not have many examples of fraudulent transactions to teach the computer. It is also very expensive to get these examples. They are often not helpful because people who commit fraud are always coming up with new ways to do it. Financial anomaly detection and fraudulent transactions are getting harder to detect because of this. Additionally, most supervised models are trained in batch mode, which limits their effectiveness in real-time streaming environments.

Unsupervised anomaly detection techniques are a way to find unusual activity by looking at what normal transactions look like and then identifying things that do not fit. This is really useful for things like banking and finance because new kinds of fraud are always popping up. Some people have already done research on using anomaly detection, with lots of financial data. This paper is going to take that research and see if we can use it with data that is coming in all the time. We are going to use something called Incremental Principal Component Analysis to make a system that can keep updating itself and find activity really quickly. The unsupervised anomaly detection techniques we are talking about can be used to make this system work well with anomaly detection. The effectiveness of the proposed approach is demonstrated using a real-world credit card transaction dataset processed in a simulated streaming environment.

LITERATURE REVIEW

Anomaly detection has been a long-standing research area in the fields of data mining and machine learning, with numerous applications in financial fraud detection. Early approaches to fraud detection were primarily based on rule-

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

based systems and statistical methods, which relied on predefined thresholds and expert knowledge. While these methods were effective for detecting known fraud patterns, they lacked adaptability and performed poorly when applied to complex, high-dimensional datasets.

With the increasing availability of large-scale financial data, machine learning-based techniques became more prevalent. Supervised learning approaches, including decision trees, support vector machines, and neural networks, have been widely used for fraud detection. Despite achieving high accuracy in controlled settings, these models depend heavily on labeled data and are sensitive to changes in transaction behavior. Moreover, the extreme imbalance between fraudulent and legitimate transactions often leads to biased models that fail to generalize well in real-world environments.

To overcome these limitations, unsupervised anomaly detection methods have gained significant attention. Isolation Forest is a widely used unsupervised technique that isolates anomalies by recursively partitioning the data space. Several studies have demonstrated its effectiveness in detecting financial fraud, particularly in high-dimensional settings. However, Isolation Forest is typically applied in batch mode and requires substantial computational resources, making it less suitable for real-time streaming applications.

Dimensionality reduction is a way to find anomalies in data. One common method is Principal Component Analysis or PCA for short. Principal Component Analysis has been studied a lot, for finding anomalies. It works by checking how well the data fits into a model of normal behavior. If the data does not fit well it is seen as an anomaly.

To make Principal Component Analysis work with sets of data and to make it fast some new versions have been made. These new versions of Principal Component Analysis can learn from data as it comes in and they do not use too much computer power. This means they can be used in time. Principal Component Analysis is still used in these versions but it is made to work with streaming data. Recent research highlights that incremental PCA-based approaches offer a favorable balance between detection performance and efficiency. Motivated by these findings, this work applies Incremental PCA for real-time anomaly detection in high-dimensional financial data streams and evaluates its performance in comparison with a batch-based Isolation Forest baseline.

PROBLEM STATEMENT

The increasing reliance on digital financial transactions has resulted in the continuous generation of large-scale, high-dimensional data streams. Financial institutions must monitor these transaction streams in real time to identify anomalous or fraudulent activities. However, designing an effective real-time anomaly detection system for financial data remains a challenging problem due to several inherent limitations of existing approaches.

Firstly, real-world financial datasets are characterized by extreme class imbalance, where fraudulent transactions constitute only a small fraction of the overall data. This imbalance significantly affects the performance of traditional supervised learning models, which tend to be biased toward the majority class. Secondly, the availability of labeled fraud data is limited, as labeling requires expert intervention and often lags behind real-time transaction processing. This makes supervised and semi-supervised approaches impractical for real-time deployment.

Furthermore, most existing anomaly detection techniques operate in offline or batch-processing modes and are computationally expensive when applied to high-dimensional data. Such approaches are unsuitable for real-time financial environments, where low detection latency and scalability are critical requirements. Additionally, evolving transaction patterns and concept drift further reduce the effectiveness of static models trained on historical data.

Therefore, there is a need for a lightweight, unsupervised anomaly detection framework capable of processing high-dimensional financial data streams in real time. The framework should adapt to incoming data, operate without labeled instances, and maintain high detection precision while minimizing computational overhead.

OBJECTIVES

The primary objective of this research is to develop and evaluate a real-time unsupervised anomaly detection framework suitable for high-dimensional financial data streams. The objectives are as follows:

1.To design an unsupervised anomaly detection model that does not rely on labeled transaction data and can effectively learn normal transaction behavior.

2.To implement an Incremental Principal Component Analysis (IPCA)-based approach that supports continuous model updates and enables efficient processing of streaming financial transactions.

3.To evaluate the performance of the proposed framework using appropriate metrics such as precision, recall, F1-score, accuracy, and ROC-AUC, with particular emphasis on handling class imbalance.

4.To compare the proposed real-time approach with a batch-based unsupervised baseline model, namely Isolation Forest, in terms of detection performance and computational efficiency.

5.To analyze the suitability of the proposed framework for real-time financial anomaly detection by examining detection latency and scalability.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

METHODOLOGY

The new system is supposed to find patterns in financial transactions as they happen. It looks at a lot of information about each transaction. The system needs to be able to handle a lot of data without slowing down.

First the system gets all the transaction information. Makes sure it is all in the same format. This helps the system work better and be more stable. The transactions are then put in order from oldest, to newest like they would happen in life. The system does not know which transactions are normal or not when it is learning. It only finds out later when it is being tested. This way the system can find patterns on its own without being told what to look for. Financial transaction data streams are what the system is looking at. Financial transaction data streams have a lot of information and the system needs to be able to look at all of it. The Incremental Principal Component Analysis is used to find anomalies in time. This is because the Incremental Principal Component Analysis can update the parts of the data without having to retrain the whole system with old data.

First we need some transactions to start with. We use these transactions to learn what normal transactions look like. After that we look at each transaction one by one.

For each transaction we calculate an anomaly score. This score shows how different the transaction is from what we think is normal. We do this by looking at how the transaction fits into our simple model of normal behavior, which is based on what we have learned so far, from the Incremental Principal Component Analysis. We look at transactions that have a lot of mistakes when we try to rebuild them. If these mistakes are really bad we mark these transactions as problems. This way we can control how careful we are when we look for problems.

To compare how well we are doing we use something called an Isolation Forest model. We train this model on a part of the data we have. The model looks for things that're very different from the rest by separating them in many different ways.

This model is really good at finding things that're different in data that has a lot of parts.. It is not very good for looking at data in real time because it needs to look at a lot of data at once. So we mostly use it when we are not in a hurry like when we're just looking at data, on our own. The new system is tested using measures like accuracy, precision, recall and how well it can tell the difference between good and bad transactions. We pay attention to precision, recall and how well the system can detect fraud because financial fraud data is very uneven.

We also check how fast the system can process transactions without slowing down which is important, for using the system to monitor transactions in real time. The system has to be able to handle transactions so it can be used in real-time financial monitoring systems.

EXPERIMENTAL SETUP

The experimental setup is made to see how well the proposed time unsupervised anomaly detection framework works on financial transaction data streams. This framework is for finding patterns in big sets of financial data. We test the proposed Incremental PCA-based approach and the baseline Isolation Forest model in a controlled environment. This way we can make a comparison between the proposed Incremental PCA-based approach and the baseline Isolation Forest model. We do all the experiments, in the conditions so the results are reliable and can be repeated. The proposed time unsupervised anomaly detection framework is what we are focusing on. This dataset has 284,807 transactions from cardholders. It has 31 things that describe each transaction like numbers and a label that says if a transaction is fake or real. The people who made the dataset had to hide the information about the transactions so they used Principal Component Analysis to make it secret. This made the dataset very big and hard to understand. It still looks like real transactions. The Credit Card Fraud Detection dataset is good, for testing because it has statistics. In dataset class are imbalance. Fraudulent transactions are very rare they make up about 0.17 percent of the total data. This makes it really hard to detect anomalies. It is a situation for testing unsupervised learning.

To make it feel like life transactions are put in order by time and treated like a stream of data that keeps coming in. Each transaction is processed one at a time not all at once so it is like a flow of fraudulent transactions and normal transactions and we have to deal with them one, by one in order to see what the fraudulent transactions look like. I used Python to do all the experiments. Python is a language for this kind of work. I also used some libraries like NumPy and Pandas to help with the machine learning and data analysis. I used Scikit-learn and Matplotlib too. I did all the experiments on a computer without any special hardware or extra power because that is how it would be in the real world.

For the Incremental PCA model that I proposed I started with a bunch of transactions to get the model going and to figure out what normal transactions look like. After that I looked at each transaction one, by one in real time using the Incremental PCA model. The number of components is chosen to keep most of the variance in the data while keeping the computer work low. Anomaly detection is done by looking at how the data can be put back together and a threshold is used to find transactions that are not normal.

The Isolation Forest model is trained on a part of the data and the settings are adjusted to make sure it can find anomalies correctly and do it quickly. Then it is tested on the data to see how well it works compared to the other method. The Isolation Forest model and the principal components method are used to find transactions. We do an evaluation to see how

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

well our system works. We use things like accuracy and precision to measure this. We also look at recall, F1-score and something called Receiver Operating Characteristic Area Under the Curve or ROC-AUC for short. We care a lot about precision, recall and ROC-AUC because the data we are working with is not balanced. This means we have a lot normal cases, than fraud cases.

We do not use class labels when we are training our system. We only use them when we are checking to see how well our system can detect fraud. We also do some checks to see if our system is strong can handle a lot of work and can work in real time. We want to make sure our system can really work with fraud detection data.

RESULTS AND DISCUSSION

Quantitative Performance Evaluation

The Incremental PCA-based anomaly detection framework is tested to see how well it works. We use accuracy, precision, recall, F1-score and ROC-AUC metrics to check this.

The dataset is not balanced so we do not just look at accuracy.

We care more about precision, recall and F1-score.

The Incremental PCA-based anomaly detection framework does a job with precision. This means it is good at not sending out alarms.

This is very important, for systems that work in real time with the Incremental PCA-based anomaly detection framework.

```
Accuracy: 0.99744
Precision: 0.7
Recall: 0.23648648648648649
F1-score: 0.35353535353535354
```

Fig. 1. Performance metrics of the proposed anomaly detection framework

Figure 1 shows the results of the evaluation metrics for the proposed framework. The proposed framework gets things most of the time which is why the accuracy is so high. The proposed framework is also very good, at finding the anomalies and it does not give a lot of false alarms, which is shown by the precision and the F1-score of the proposed framework.

The confusion matrix gives us an idea of how the proposed model works when it comes to classifying things. It shows us which transactions were classified correctly and which ones were not. The proposed model is what we are looking at here so we want to see how it does with the proposed models classification. This helps us understand the proposed model better.

Confusion Matrix Analysis

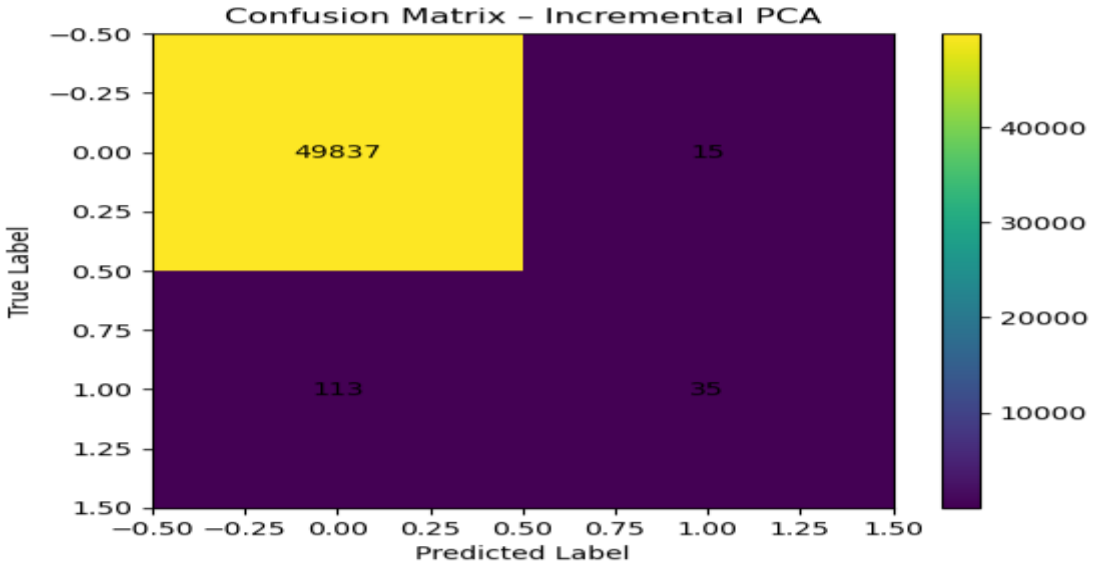

Fig. 2. Confusion matrix for the Incremental PCA-based anomaly detection model

If you look at Figure 2 you can see that the model gets it right when it comes to transactions. It says they are normal which is what we want. The model does not send out a lot of alerts, which is good. This means it does not bother us with warnings that're not necessary. The model does miss some transactions but that is okay. In the world when we are dealing with money

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

it is very important that we do not get too many false warnings. This is because we want to make sure we are not stopping transactions by mistake. The model is being careful. That is what we need for financial things that happen in real time.

Anomaly Score Distribution Analysis

Anomaly detection with PCA is really about finding mistakes. It looks at how wrong the transactionsre from what we think is normal. We do this by measuring the reconstruction error. This error shows us how different the transactions are from what the Incremental PCA has learned is behavior. Anomaly detection using Incremental PCA is, about this reconstruction error.

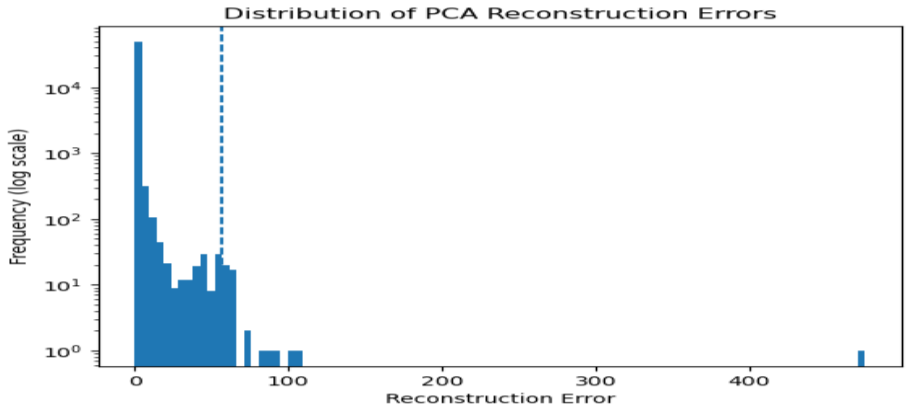

Fig. 3. Distribution of PCA reconstruction errors with anomaly threshold

Figure 3 shows us how the mistakes in rebuilding transactions are spread out for every transaction. A lot of transactions have small mistakes when we try to rebuild them which means they are working normally.. Some transactions have really big mistakes. The transactions that have mistakes than a certain limit are considered unusual. This shows that we can clearly tell the difference, between transactions and unusual transactions.

Temporal Analysis of Anomalies

The proposed framework is tested to see how well it can detect things in time. To do this the framework gives each transaction an anomaly score. These scores are then plotted on a graph as the transactions happen one, after the other. This shows how the framework does over time as it looks at each transaction. The anomaly scores of the proposed framework are what is being looked at here.

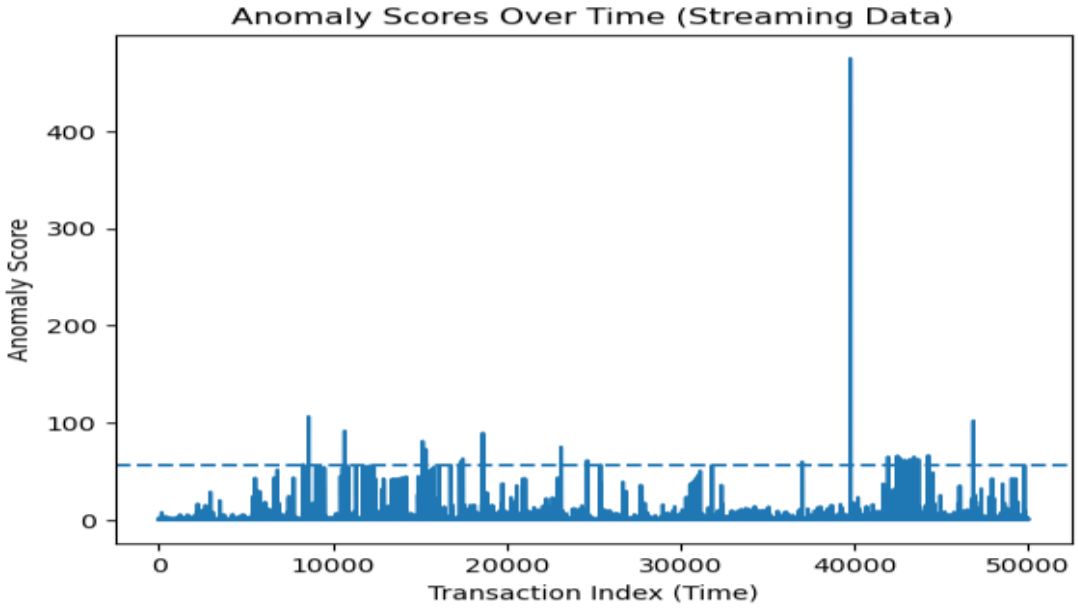

Fig. 4. Anomaly scores over time in streaming financial transactions

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

If you look at Figure 4 you can see that the anomaly scores are usually low for transactions.. Sometimes there are big jumps that go above the threshold for detection. These big jumps are a sign of transactions that're not normal or could be fraudulent. This shows that the method we are proposing can find anomalies as they happen in time with the anomaly scores and the detection threshold for the transactions. The anomaly scores are important for the transactions. Help with the detection of anomalies, in real time.

ROC Curve Analysis

The Receiver Operating Characteristic curve is a way to see how well the anomaly detection framework can tell the difference between things that're normal and things that are not. The Receiver Operating Characteristic curve is really good, for figuring out if the anomaly detection framework is working like it should. We use the Receiver Operating Characteristic curve to evaluate the anomaly detection framework.

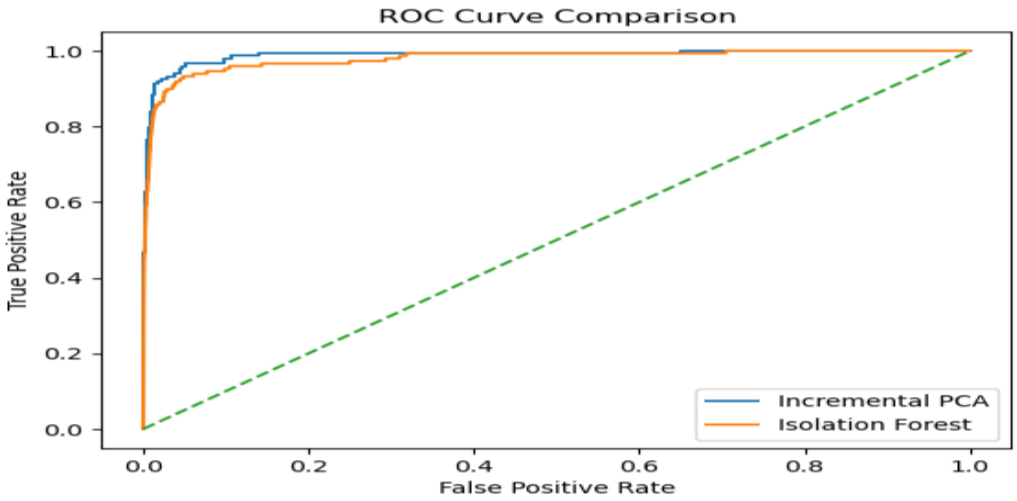

Fig. 5. ROC curve comparison between Incremental PCA and Isolation Forest

Figure 5 shows that the Incremental PCA-based model does a job. It gets a ROC-AUC value. This means the Incremental PCA-based model is really good at telling transactions and anomalous transactions apart.

When we compare the PCA-based model to the Isolation Forest baseline we see that it works just as well at finding anomalies. The Incremental PCA-based model is also very good because it does all of this in time.. It does not need as much computer power, as other methods. This is a deal because the Incremental PCA-based model can keep up with a lot of transactions without slowing down.

Discussion Summary

The Incremental PCA-based framework really works well. It does a job of finding a balance between being accurate and not using too much computer power. The model is very precise. It works well when it has to make decisions quickly. It can even do this in time when it is looking at a stream of data.

The Incremental PCA-based framework is not as good at finding all of the problems. This is okay. This is because it is better to be safe than sorry especially when it comes to money. The Incremental PCA-based framework is good at not sending out alarms. This is very important for applications.

Overall the results show that the Incremental PCA-based framework is a choice for finding unusual activity in real time in financial data that has a lot of information. The Incremental PCA-based framework is a tool, for this job.

CONCLUSION

This paper is a way to find unusual activity in financial data as it happens. It uses a method called Incremental Principal Component Analysis to understand what normal transactions look like and then find transactions that do not fit this pattern. The financial data is looked at one piece at a time not all once and it does not need to be labeled beforehand. This makes it very useful for institutions where things are happening quickly and they need to make decisions fast. The Incremental Principal Component Analysis method is used to model transaction behavior and then it finds anomalies based on how well it can rebuild the transactions. This is different from methods that need a lot of data to be labeled and then looked at all, at once. The framework is updated as new transactions happen, which makes it work well in time financial environments where financial data streams are high-dimensional.

We tried out the model using a real credit card transaction dataset from the real world. This dataset had a problem. It had a lot of normal transactions and very few fraudulent ones. The results showed that our new model worked well. It was very

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

good at finding the fraudulent transactions and it did not make many mistakes. We also looked at how the model performed in terms of positives, which is a big deal for financial fraud detection systems. We do not want the model to say something is fraudulent when it is not. The model did a job of keeping these mistakes to a minimum, which is what we want for a real-world credit card transaction system, like this. The recall is not as good because of the limits that are, in place.. This is okay when we are talking about things that need to happen in real time. In these situations it is more important to avoid alarms. This is what matters most when we are dealing with real-time applications and the recall of these applications.

The findings show that using methods that do not need supervision and can learn from new data can really help find anomalies in big financial data that is always coming in. The framework that is proposed is good because it balances finding anomalies being able to handle a lot of data and being able to work in real time which makes it a good solution for financial monitoring systems that are used in the real world. Lightweight methods for finding anomalies are very useful, for data.

FUTURE SCOPE

The proposed framework is really good at finding anomalies in time.. There are some things we can do to make it even better. One thing we can try is adding a feature that automatically changes the threshold for what we consider an anomaly. This means the framework will look at how people're using their money and change what it thinks is weird based on that. This will help the framework find anomalies without getting too many wrong answers, which is important when people are using their money in different ways. The framework will be able to do this when the way people use their money is changing, like, in the financial world.

Future work may also look at how to add tools that can find and adjust to changes, in the way people make transactions over time. This will make the model work better when it is used in the world for a long time. We can also try combining PCA with other models that do not need supervision or only need a little supervision to see if it can find things better without using too much computer power. Incremental PCA can be used with models to make it work even better.

The financial data framework can be made better by using it with different sources of financial information. This includes things like details, about transactions, records of what people do and patterns that happen over time. The financial data framework can be tested with big sets of financial information. It can also be used in financial systems to see how well it works when it is actually being used. This will show if the financial data framework is effective when it has to deal with world limits. The financial data framework is what needs to be evaluated and used in these situations.

REFERENCES

[1] 1.F. T. Liu, K. M. Ting, and Z. H. Zhou, “Isolation Forest,” Proceedings of the IEEE International Conference on Data Mining, pp. 413–422, 2008.
[2] 2.R. Chalapathy and S. Chawla, “Deep Learning for Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 52, no. 1, pp. 1–38, 2019.
[3] 3.A. Zimek, E. Schubert, and H.-P. Kriegel, “A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data,” Statistical Analysis and Data Mining, vol. 5, no. 5, pp. 363–387, 2012.
[4] 4.C. C. Aggarwal, Outlier Analysis, 2nd ed., Springer, 2017.
[5] 5.H. M. Gomes, A. Bifet, J. Read, et al., “Adaptive Random Forests for Evolving Data Stream Classification,” Machine Learning, vol. 106, pp. 1469–1495, 2017.
[6] 6.U. Dal Pozzolo, O. Bontempi, and G. Snoeck, “Adversarial Drift Detection for Fraud Detection,” IEEE Intelligent Systems, vol. 30, no. 3, pp. 15–20, 2015.
[7] 7. D. Dal Pozzolo and his team, including G. Bontempi and O. Snoeck wrote a paper called "Calibrating Probability with Undersampling for Unbalanced Classification". This paper was presented at the IEEE Symposium on Computational Intelligence and Data Mining. The paper is, on pages 159 to 166. It was published in 2015. D. Dal Pozzolo and his team did a job with this paper on Calibrating Probability with Undersampling for Unbalanced Classification.
[8] 8.J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, “A Survey on Concept Drift Adaptation,” ACM Computing Surveys, vol. 46, no. 4, pp. 1–37, 2014.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)