Synthetic Data Generation:

Paper Link : https://www.isij.eu/system/files/download-count/2024-11/5534_Fraud_detection.pdf

Inferences:

1. Problem with all the ML methods was coming as whatever data we had, fraud instances were less so model didnt learn those properly. Isliye we need to have better data which incorporates both fraud and not fraud wali stuff together.
2. There was SMOTE but that created unrealistic data, basically model kharab.
3. GANs are kinda nais but there is issue like mode collapse (?) and instability during training.. To logistically not so helpful
4. Then comes diffusion, basically made to tackle the problem with GAN. data is generated by refining samples through various transformations.. Gives realistic and diverse output, nais cuz then model is learning proper stuff.
5. Model that they used: Duo-GAN that is modified to employ a diffusion process.

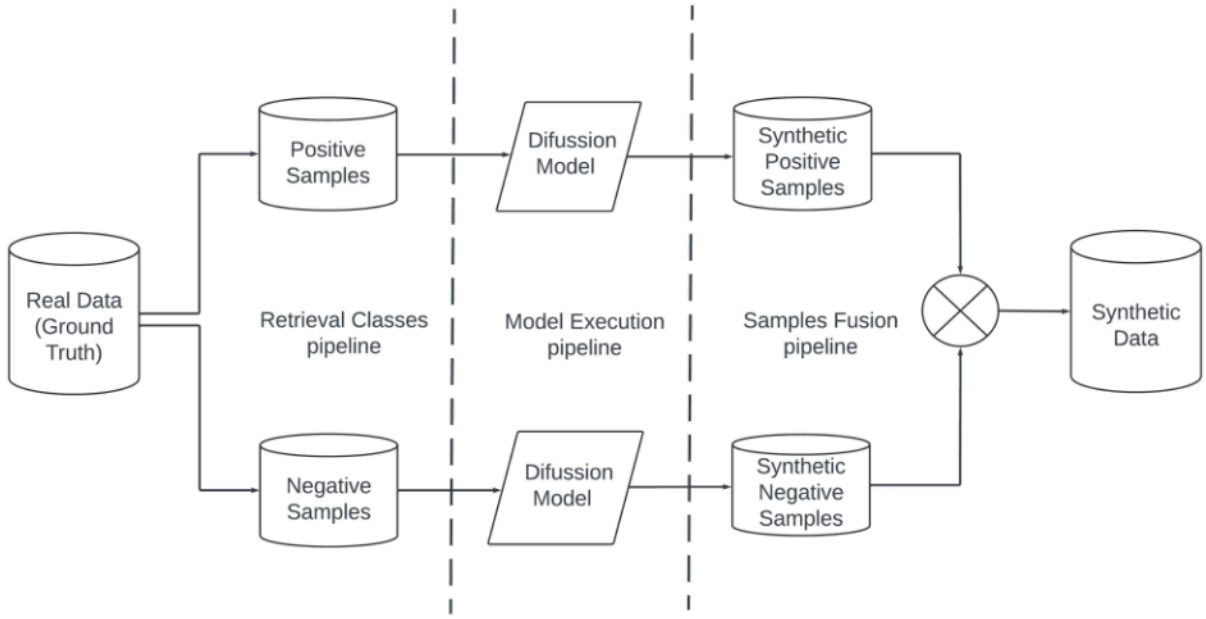


Figure 1: The blueprint of SDG using a diffusion model.

6. 2 separate diffusion blocks that learn independently from positive sample. Now they capture this ‘class-conditional distributions and correlations from each class’(# on trying to understand, seems like it basically tries to understand what makes a not fraud transaction not fraud and basically tries to establish that type of relationship along the whole data, this way jo naya data banega vo is cheez ko dhyan me rkhe hue hoga), helps in understanding of the actual data distribution. 2 tracks because ek me positive ka data jayega and ek me negative ka.

7. BTW, ALSO FOUND A DATASET THROUGH THIS WE CAN USE. -> IEEE-CIS Fraud Detection Dataset

Dataset Name	Domain	Number of features	Number of Instances	Imbalance Ratio (IR)
IEEE-CIS Fraud Detection Dataset	E-commerce	67	561,013 (training), 28,527 (testing)	1:28.6

Issues: how to preserve data privacy

Well.. https://www.jstage.jst.go.jp/article/ipsjjip/30/0/30_789/_pdf. This paper basically is like banks sending data to each other and preventing fraudulent transactions by like having ciphers generated through SGD and every back having to decrypt this cipher. All the while, individual data is not shared, basically only sharing statistical data. This allows us to get large amounts of data legally.

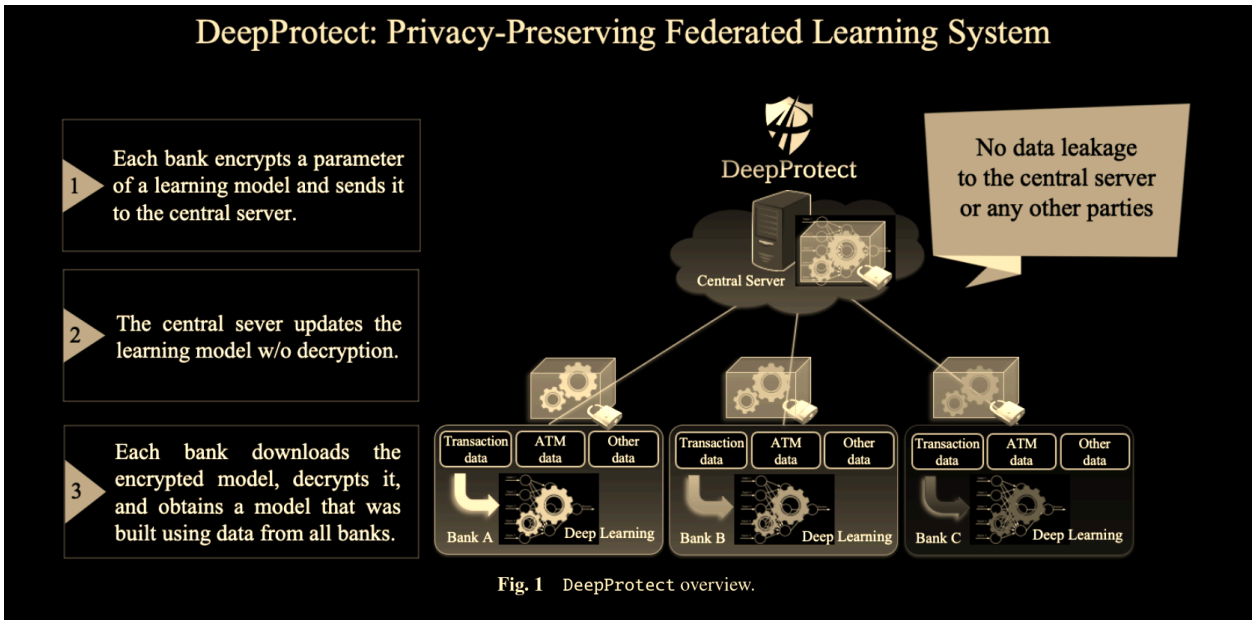
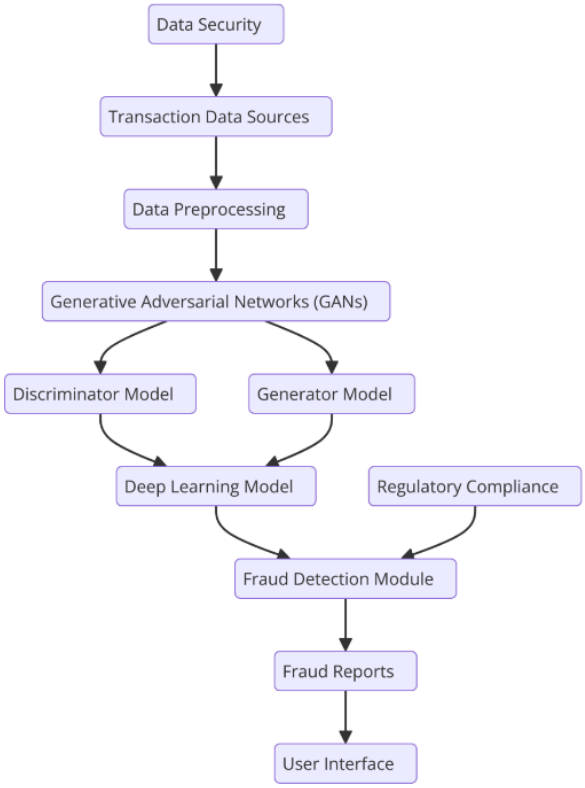


Fig. 1 DeepProtect overview.

This is a crazy good paper as it basically leverages banks to _sort of _ share data of transactions for better modelling and like an 80%+ fraud detection rate

Problems: Speed and BlackBox-ish nature, regulatory compliance needed

Therefore: https://thesciencebrigade.com/JAIR/article/download/402/376/828

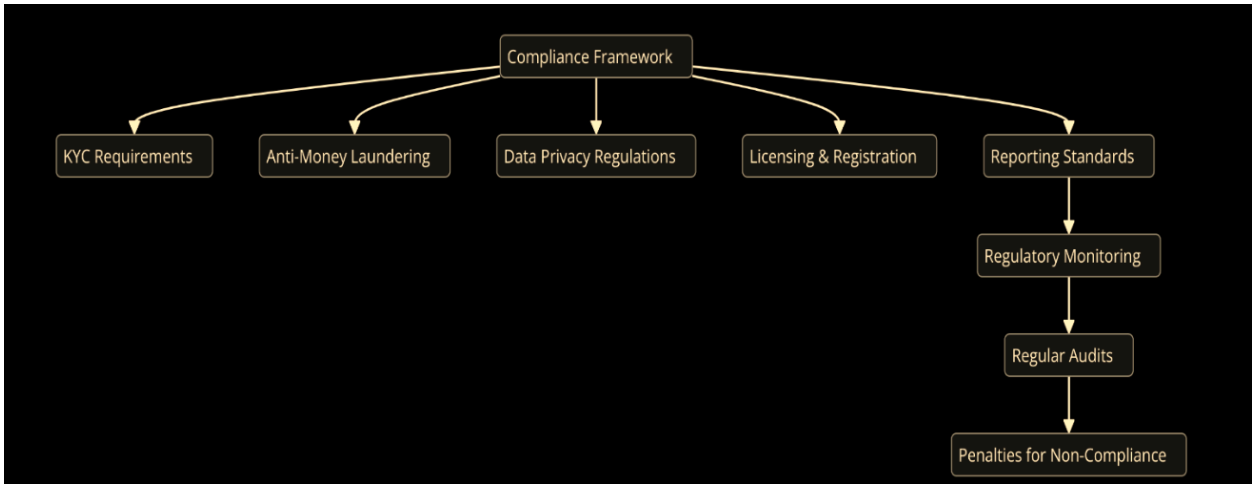


In the DL model, there is basically CNN and RNNs(specifically LSTMs). The GAN creates crazy data while doing the classification stuff and CNN and LSTMs learn the spatial and temporal patterns in the data.. Final layer works on risk scores, high risk score is flagged and rest are passed forward to do normal payments.

In the data taking stage, its basically taking crazy nuances in as well.. Like how was the transaction initiated and stuff

transaction sequences and user activity logs are converted into format suitable for LSTM. GNN matches the incoming data with the synthetic one it creates, CNNs try to see if say crazy transactions happen in like the similar region and LSTM sees the frequence..

The risk score is basically the output from the DL model.



Compliance stuff.. Vaise to GAN n all help in the “report irregularity ASAP” part.. GDPR(privacy part) im thinking to incorporate the homologous encryption that i mentioned earlier.

Implementing methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) could enhance transparency and provide stakeholders with insights into model predictions.

—

Found this https://github.com/illidanlab/HyFL/tree/main?tab=readme-ov-file, its like an improvement to the above shared federation model but genwinly kuch zyada dhang se samajh nahi aa rha… trying hard.
(https://arxiv.org/pdf/2302.03654 ← corresponding research paper)