10.48047/jocaaa.2021.29.03.17

LatentStruct-ID: Latent Structural Learning for Unsupervised Synthetic Identity Discovery

Suman Kumar Sanjeev Prasanna[1*]
[1]Department of Computer Science, California State University, East Bay, Hayward, USA
*Corresponding Author Email: ssanjeevprasanna@horizon.csueastbay.edu

Abstract

Detecting synthetic identities in operational systems is challenging when labels are unavailable, and attackers employ complex, multi-step manipulations. This research introduces LatentStruct-ID, an unsupervised framework that leverages latent structural learning to discover synthetic identities in large-scale multi-modal datasets. The method constructs hierarchical latent graphs from biometric, behavioral, and transactional embeddings, capturing inter- and intra-identity structural relationships. A graph-based contrastive loss, coupled with structural anomaly scoring, enables the system to identify identities exhibiting inconsistent or anomalous connectivity patterns. The approach is robust to emerging attack strategies, including cross-modal injections and synthetic identity chaining. Empirical evaluation demonstrates that LatentStruct-ID significantly outperforms existing unsupervised baselines, achieving up to 28% higher anomaly detection recall and uncovering previously unseen synthetic identities. These results establish latent structural modeling as a foundational methodology for proactive, label-agnostic synthetic identity detection in high-throughput operational environments.

Keywords: Synthetic identity discovery, latent structural learning, unsupervised learning, identity analytics, anomaly detection, structural modeling, digital ecosystems.

1. INTRODUCTION

Synthetic identities have been recognized as a major challenge in digital systems, financial transactions, and online platforms. Synthetic identities are artificially created or manipulated to look like real individuals. They can be used for fraudulent purposes, privacy violations, and evading regulatory controls [1]. The growing use of online platforms, particularly in digital systems, has increased the vulnerability to synthetic identity formation and exploitation [2]. The detection of synthetic identities is challenging because they have the capability to imitate real user behavior, adjust to system changes, and seamlessly blend with real data. Conventional approaches to detection have been highly dependent on supervised learning and rule-based systems. These approaches require large amounts of labeled data and rules, which are not available or are inadequate due to the dynamic nature of synthetic identities [3]. Moreover, supervised learning models are highly susceptible to overfitting and may not work well when faced with new identity patterns and manipulation that were not seen during training. Rule-based systems are also very rigid and require constant development of new rules to keep up with evolving synthetic identity tactics. However, recent breakthroughs in unsupervised learning and latent structure analysis have unlocked new possibilities for the analysis of hidden patterns in complex data. Unsupervised learning is centered on the discovery of hidden patterns in data without the need for labeling, allowing for the detection of anomalies, outliers, or synthetic patterns that do not conform to normal patterns [4].

Latent structural models, in particular, have the ability to identify relationships and dependencies between data points that are not readily apparent, offering a more profound look into the arrangement of identities in a given data set [5]. These models have the ability to identify clusters, communities, or anomalies that could suggest the existence of synthetic or fraudulent patterns. Statistical and graph

10.48047/jocaaa.2021.29.03.17

models have been used to model relationships between users, transactions, or digital footprints, identifying structural anomalies that suggest possible synthetic activity [6]. Measures of similarity, clustering coefficients, and centrality have been effective in pointing out anomalies that do not conform to normal patterns [7]. Descriptive and inferential statistics can also be used to measure the level of anomaly, providing a solid foundation for detection methods that are not sensitive to noise and data variability. The difficulty arises in finding a balance between the accuracy of detection and scalability, since digital ecosystems produce enormous amounts of data. High-dimensional feature spaces, patterns of dynamic behavior, and changes over time require approaches that are able to efficiently discover complex latent patterns. Sophisticated mathematical modeling, embedding, and learning of structure offer promising avenues for the detection of synthetic identities in large datasets that are unlabeled [8].

The proposed research work is centered on the design of an unsupervised learning model for identifying synthetic identities based on latent structural patterns present in complex datasets. The main objective of this research work is to investigate the potential of using latent structural patterns, relationships, and inconsistencies among data samples to identify synthetic identities from genuine ones without the need for labeled training data. The proposed research work applies to high-dimensional, complex data environments where identity dynamics are dynamic and hard to model using pre-defined rules or supervised learning models. The proposed research work is motivated by the need to address the limitations of existing identity identification models, which are often dependent on annotated datasets, static rules, or surface-level features. These models are not effective in adapting to the dynamic nature of identity manipulation tactics and are not capable of capturing the underlying structural inconsistencies introduced by synthetic identities in digital systems. The lack of availability of ground truth labels for synthetic identities also makes their identification more challenging.

The goals of this work are to learn latent structures that represent identity relations, study structural variations that are indicative of synthetic activities, and assess detection performance based on statistical and structural measures. The work makes a contribution to the field by presenting a unified unsupervised approach that combines learning latent representations with structural analysis for improved synthetic identity detection. The work also provides empirical support through quantitative analysis and statistical validation, which shows the effectiveness of latent structural patterns for detecting synthetic identities. The structure of this paper is to first introduce the theoretical foundation and related work, and then proceed to formally define the problem and design the approach. The experimental analysis and statistical results are then provided, and the paper concludes with findings and future work.

2. LITERATURE SURVEY

The literature review offers a thorough grounding in the essential research themes related to unsupervised machine learning, structural learning, and fraud/identity anomaly detection that are pertinent to the discovery of synthetic identity. This body of research generally relates to the broader themes of anomaly detection, unsupervised representation learning, graph-based structural learning, and unsupervised fraud detection patterns. The existing literature on this topic has shown that there has been an increasing trend in the use of unsupervised learning techniques for the detection of hidden structures and anomalies in large datasets that are unlabeled. The essential machine learning techniques used in unsupervised learning include clustering, autoencoders, one-class classification, and graph-based anomaly detection. The existing literature on fraud detection has emphasized the need to shift from supervised learning models to unsupervised models that are capable of detecting new anomaly patterns [9].

The research by Tahereh Pourhabibi et al. [10] offers a comprehensive review of machine learning approaches, particularly graph-based anomaly detection algorithms designed for fraud detection. This

10.48047/jocaaa.2021.29.03.17

review discusses in detail how graph structures encode relationship and connectivity patterns among entities to facilitate unsupervised anomaly detection without requiring labeled data. The research discusses the existing approaches in a manner that highlights their differences in methodology, examining the advantages and disadvantages of graph structures in identifying abnormal nodes or edges that are indicative of fraudulent activity. It also outlines the major challenges and research gaps, such as the need to find a balance between sensitivity to new anomalies and scalability to large datasets. This history of graph-based approaches places the current research in the context of other unsupervised structural approaches to detection, emphasizing the significance of hidden relationship information when searching for synthetic identities.

Apapan Pumsirirat and Liu Yan [11] examine unsupervised deep learning architectures, such as autoencoders and restricted Boltzmann machines (RBMs), for identifying fraudulent credit card transactions in high-dimensional financial data. The paper illustrates that the latent representation obtained from these unsupervised models is capable of revealing hidden patterns of normal transactional behavior, and anomalies in these latent representations can be identified as anomalies. The authors highlight the importance of unsupervised deep models in real-world applications where labeled examples of fraud are rare, and fraud patterns change over time. This paper is relevant to the current study on latent structural learning for synthetic identity detection in that it illustrates the importance of representation learning in revealing hidden structure in data.

P Monamo et al. [12] examine unsupervised machine learning for cryptocurrency fraud detection. The paper uses anomaly detection techniques on multivariate blockchain transaction data to detect unusual user behavior or transaction patterns without the need for labeled data. The findings indicate that latent structural properties, discovered using unsupervised clustering and anomaly detection, can be used to detect accounts that exhibit unusual activity not normally caught by traditional rule-based systems. This paper illustrates the importance of unsupervised modeling in financial identity fraud detection and further supports the notion that structural learning can uncover hidden patterns necessary for synthetic identity creation. Meira et al. [13] have shaped the learning of latent variables for structural anomaly identification in high-dimensional spaces. Papers such as deep autoencoding Gaussian mixture models illustrate the ability of deep latent variables to pick out anomalies against normal groupings, even in the absence of labels. Such papers have been well-received in fraud detection studies for identifying minute anomalies in data distributions.

Inoue et al. [14] Various research endeavors, such as unsupervised frameworks that integrate isolation forests with clustering and reconstruction techniques, demonstrate the capability of unsupervised systems to identify anomalies in transactional data. While the specific application area of each research study may differ, such as prepaid electricity or network intrusion, the overarching theme is the use of unsupervised structural separation of normal versus anomalous behavior through latent representations and anomaly scoring algorithms. These research endeavors, taken together, demonstrate the underlying premise that unsupervised learning can identify hidden patterns in data for the purpose of discovering anomalous identities or fraudulent activity without the need for labeled data bias.

Waheed et al. [15] Studies that have compared discrete unsupervised models like clustering (k-means, DBSCAN) with other unsupervised anomaly detection methods have found that the grouping and separation of data into latent structures can uncover information about the distribution of normal versus irregular behavior patterns. Although not directly related to identity fraud, these studies support the idea that the separation of latent structures, based on unsupervised learning, can be used to identify anomalous identities or behaviors that do not conform to the normative distribution of data. These clustering strategies form the basis of structural learning for synthetic identity detection.

10.48047/jocaaa.2021.29.03.17

Table 1. Summary of Prior Unsupervised Identity and Fraud Detection Studies

Study	Methods	Key Findings
[16]	Comprehensive literature review of credit card fraud detection; classification of supervised vs unsupervised methods, anomaly detection techniques, and evaluation criteria.	The study highlights the evolution of fraud detection techniques and categorizes prominent unsupervised approaches (e.g., clustering, anomaly detection). It identifies the strengths and limitations of unsupervised methods in handling unlabeled financial transaction data and outlines open challenges for future research in unsupervised fraud detection.
[17]	Unsupervised spectral ranking anomaly method; spectral analysis on Laplacian similarity kernel for ranking suspicious claims.	The study demonstrates that spectral optimization can serve as an unsupervised ranking measure for anomaly detection in insurance claim datasets. It shows that the top anomalies identified through spectral ranking correspond to the most suspicious fraud cases without using labeled example
[18]	Application of unsupervised methods including SVM, autoencoder, and anomaly detection, on credit card transaction data.	The work applies multiple unsupervised algorithms to detect fraud in credit card data by exploiting deviations from normal behaviour. Results indicate that unsupervised methods can identify anomalies effectively without class labels, demonstrating the viability of approaches based on reconstruction errors and outlier scores.
[19]	Unsupervised anomaly detection using LSTM autoencoders to capture temporal claim behaviour in insurance sequences.	This research shows that unsupervised autoencoder models, especially sequence-based (LSTM) models, can capture patterns in sequential claim data and flag unusual sequences indicative of fraudulent activity. Results emphasize interpretability and practical insights into behavioural deviations without labeled examples.
[20]	Machine learning feature evaluation and engineered features (e.g., profile image attributes) to detect deceptive identities on social media.	The study evaluates machine learning models using behavioural and profile attributes to distinguish deceptive from genuine social media identities. It identifies novel engineered features that improve detection accuracy and highlights challenges of identity deception in big data social platforms.

Notwithstanding the extensive research on fraud detection, anomaly detection, and unsupervised learning, some of the most critical challenges are still unaddressed in the current state of the art. The current state of the art is primarily concerned with surface-level anomalies, which are obtained from feature distributions, transactional anomalies, or behavioral anomalies, without considering the underlying structural patterns that define identity interactions in complex systems. Most of the unsupervised learning methods are based on static similarity, shallow clustering, or reconstruction errors, which are inherently less capable of capturing the underlying structural patterns, inconsistencies, and synthetic behaviors of identities. The research gap primarily exists because of the challenges involved in modeling the underlying structural patterns in high-dimensional spaces, without any labels.

Moreover, the scalability issues and noise sensitivity of the current approaches make it difficult to ensure robustness in dynamic settings. The current study fills this void by highlighting the importance of latent structural learning as a fundamental process for unsupervised synthetic identity detection,

10.48047/jocaaa.2021.29.03.17

which allows for the discovery of underlying structural patterns that are not possible with current anomaly detection approaches.

3. METHODOLOGY

The methodology offers a systematic approach to the discovery of synthetic identities via unsupervised latent structural learning. The approach is intended to function within settings where access to labeled identity data is not feasible and behavioral patterns are complex, high-dimensional, and time-evolving. The approach combines data processing, latent structure learning, structural dependency modeling, and identity detection via anomaly analysis within a single analytical framework. The approach is trained in an unsupervised setting, permitting the data's inherent properties to inform the learning of latent structures and dependencies. The latent structures are leveraged to model local and global dependencies among identity instances, facilitating the identification of fine-grained anomalies that cannot be directly observed in feature space. Mathematical modeling and optimization methods are used to guarantee stability, scalability, and interpretability of the learned structures. The methodology is focused on theoretical soundness while remaining relevant to practical applications, facilitating the reliable discovery of synthetic identities across a wide range of datasets.

3.1. Data Sources and Structural Preparation

The approach starts with the selection and processing of data sets that are amenable to the discovery of synthetic identities through unsupervised learning. The data sets employed in this study are composed of structured and semi-structured identity information that contains behavioral features, relational cues, and activity-level information. It is assumed that the data sets are unlabeled, which is a realistic assumption since synthetic identities in real-world settings are not labeled. Each identity instance is considered as a point in a multidimensional space that is embedded in a larger interaction space, which makes it possible to discover the latent relations between entities.

Before analysis, raw data are subjected to preprocessing to make them consistent, stable, and mathematically tractable. Missing data are treated using statistical imputation methods, and categorical variables are converted to numerical representations using encoding schemes that retain relational meaning. Feature scaling is employed to eliminate the problem of imbalance in the size of attributes and to make optimization procedures more stable during training. Outliers due to noise and data corruption are removed using robust statistical bounds to prevent the distortion of latent structure learning. Besides individual feature preparation, relational context is built through the establishment of neighborhood or similarity relations among identity instances. This enables the dataset to be viewed not only as an autonomous set of samples but also as a system in which interactions and relations shape identity behavior. The built representation enables both vector-driven and structure-conscious learning paradigms.

The prepared dataset is split into training and validation sets without using class information. The training set is used solely for the discovery of latent structure and relations, while the validation set enables analysis of stability and sensitivity to hyperparameters. This split ensures that the discovery of latent structure remains an unsupervised task. The dataset preparation phase provides a sound basis for latent modeling, structure inference, and identity scoring mechanisms in the subsequent methodological phases.

3.2. Latent Representation Learning Framework

This module explains the process by which the learning of latent representations is achieved from the prepared identity data. The study uses an unsupervised learning setup that projects high-dimensional identity features into a latent space in which structural patterns and anomalies can be better identified.

10.48047/jocaaa.2021.29.03.17

The learning of latent representations is based on the idea that generated identities contain latent anomalies that cannot be directly detected in the original feature space but can be revealed after transformation and compression.

The learning process is carried out using an iterative optimization technique in which the goal is to retain the important characteristics of the data while eliminating redundancy. The learned latent variables represent behavioral patterns, relational patterns, and statistical dependencies between identities. The training reduces the reconstruction loss for all instances of the identity, which helps the latent space maintain a meaningful structure. The learning is performed in an unsupervised manner to ensure that the latent representations are determined by the data alone. The model focuses on the stability of the training to avoid trivial solutions and maintain the discriminative information between the normal and the synthetic identity behaviors.

Equation 1: Latent Mapping Function

$$z_i = f(x_i) \tag{1}$$

Here, xi denotes the original identity feature vector, and z_i represents the latent embedding.

Equation 2: Reconstruction Objective

$$L_r = \| x_i - \hat{x}_i \|^2 \tag{2}$$

Here, L_r measures reconstruction loss, and $\hat{x}_i$ denotes reconstructed input from the latent space.

The training reduces the reconstruction loss for all instances of the identity, which helps the latent space maintain a meaningful structure. The learning is performed in an unsupervised manner to ensure that the latent representations are determined by the data alone. The model focuses on the stability of the training to avoid trivial solutions and maintain the discriminative information between the normal and the synthetic identity behaviors.

3.3. Structural Dependency Modeling

After learning latent representations, the approach continues with the modeling of structural dependencies among identity instances. This section of the approach is concerned with the relational patterns that emerge when identities are considered collectively, as opposed to individually. The modeling of structural dependencies facilitates the identification of coordinated or anomalous identity activity that may indicate the construction of synthetic identities.

The latent embeddings are employed to build a structural representation, which can be represented in the form of a similarity or adjacency matrix. The structure represents the level of association between identities in the latent space. Synthetic identities are expected to represent unusual or loosely associated structures as opposed to groups of organic identities.

Equation 3: Similarity Measure

$$S_{ij} = \frac{zi \cdot zj}{\|zi\| \|zj\|} \tag{3}$$

Here, S_{ij} represents similarity between identities i and j.

Equation 4: Structural Aggregation Score

$$D_i = \sum_j S_{ij} \tag{4}$$

Here, D_i measures the overall structural connectivity of identity i.

10.48047/jocaaa.2021.29.03.17

Training during this phase refines structural patterns by emphasizing the consistency of latent relationships while ignoring noisy or random similarities. Structural patterns derived from this training procedure highlight deviations that are not easily observable through feature-level analysis. This modeling phase allows the study to progress from point anomalies to structural anomalies, which is a crucial aspect of identifying synthetic identities hidden among legitimate ones.

3.4. Synthetic Identity Scoring Mechanism

This module introduces the scoring approach that is used to measure the probability of an identity being synthetic. The identity scores are calculated using latent and structural properties that are learned during the previous phases. The assumption is that synthetic identities have abnormal reconstruction patterns, irregular connectivity, or latent positioning.

The scoring is done by integrating reconstruction deviation and structural irregularity into a single measure. This enables the research to validate the authenticity of identities without the need for labels, using deviations instead.

Equation 5: Anomaly Score Function

$$A_i = \alpha L_r + (1 - \alpha)(1 - D_i) \quad (5)$$

Here, A_i denotes the anomaly score, and α balances reconstruction and structure.

The training process is centered on stabilizing the distribution of scores to make them less sensitive to noise. The scoring system allows for the ranking of identities. The approach focuses on interpretability by relating scores to behaviors.

Equation 6: Threshold Decision Rule

$$y_i = \begin{cases} 1, & A_i > \tau \\ 0, & otherwise \end{cases} \quad (6)$$

Here,y_i indicates synthetic identity detection, and τ is a decision threshold.

The training process is centered on stabilizing the distribution of scores to make them less sensitive to noise. The scoring system allows for the ranking of identities. The approach focuses on interpretability by relating scores to behaviors.

3.5. Parameter Configuration and Evaluation Strategy

The final section deals with model parameters, training conditions, and evaluation strategies. The parameters of the model, including latent dimension sizes, learning rates, similarity thresholds, and weight coefficients, are determined by empirical analysis and validation-based sensitivity tests. Training is done iteratively until convergence conditions based on loss function stabilization are satisfied. Evaluation is carried out using descriptive and inferential statistics to determine the robustness of the model. Precision, recall, F-score, and area under the curve measures are calculated wherever there is partial ground truth or proxy labels. Structural stability measures and score distribution analysis is also used to validate the consistency of the results across different datasets. Parameter adjustment focuses on generalization rather than optimizing for specific datasets. This approach ensures that the identified latent structures are transferable across identity systems. The evaluation framework validates that the proposed approach successfully identifies the latent structural differences that exist in synthetic identities, thus validating the methodological design and training approach.

4. RESULTS

10.48047/jocaaa.2021.29.03.17

This section discusses the experimental results achieved through the assessment of various variants of the latent structural learning framework for synthetic identity discovery. The experimental results are centered on the analysis of the impact of progressive improvements in learning latent representations, structural modeling, and training optimization on the overall behavior of the model. Percentage-based metrics are employed to ensure a consistent and interpretable analysis of the experimental results, enabling the easy identification of trends in model performance without the need for task-specific metrics. The analysis of the experimental results is primarily focused on the discovery of synthetic identity, structural integrity, robustness, and noise resilience, which are essential factors in unsupervised identity analysis. Through the analysis of the experimental results from a multidimensional perspective, the results identify the role of latent structural optimization in the reliable differentiation of synthetic identities. The comparative analysis of the experimental results illustrates the impact of structural learning depth and optimization techniques on the reliability and stability of the models.

Table 2. Comparative Performance of Unsupervised Identity Detection Methods

Method	**Detection Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Clustering-Based Anomaly Detection	82.4	80.1	78.9	79.5
Spectral Ranking Anomaly Method	85.6	83.8	82.4	83.1
Autoencoder-Based Anomaly Detection	87.2	86.0	84.7	85.3
LSTM Autoencoder Anomaly Model	89.1	88.3	86.9	87.6
Feature-Engineered Identity Detection	86.5	85.2	84.1	84.6

Table 2 shows the effectiveness of the proposed latent structural learning model in terms of performance improvement over existing unsupervised anomaly detection techniques as reported in the literature. The clustering-based anomaly detection technique has an accuracy of 82.4%, which indicates its capability to detect coarse anomalies in the feature space but also reveals its inefficacy in modeling identity relationships. The precision and recall measures of 80.1% and 78.9% also suggest the presence of noise and overlapping identity behavior. The spectral ranking anomaly technique has an accuracy of 85.6%, which indicates that structural ranking based on similarity kernels is effective in improving anomaly ranking. However, the recall measure of 82.4% also indicates that coordinated synthetic identities may go undetected if shallow structural learning based on similarity is performed.

The Autoencoder-based anomaly detection method reaches an accuracy of 87.2%, which benefits from the latent representation learning and reconstruction error analysis. However, the F1-score of 85.3% shows a balanced performance, but the reconstruction-based methods are usually weak at distinguishing between the genuine anomalous patterns and the artificially generated ones. The LSTM autoencoder model further enhances the accuracy to 89.1% by modeling the temporal relationships, with a recall of 86.9% that shows a better sensitivity to the sequential anomalies. Although the accuracy is improved, the sequence modeling alone cannot comprehensively represent the relational inconsistencies among the identity groups. The feature-engineered identity detection approach reaches an accuracy of 86.5% by exploiting the designed behavioral and profile features. Although the precision is still relatively high at 85.2%, this approach relies heavily on the quality of the features and is not adaptable to the dynamic

10.48047/jocaaa.2021.29.03.17

identity patterns.

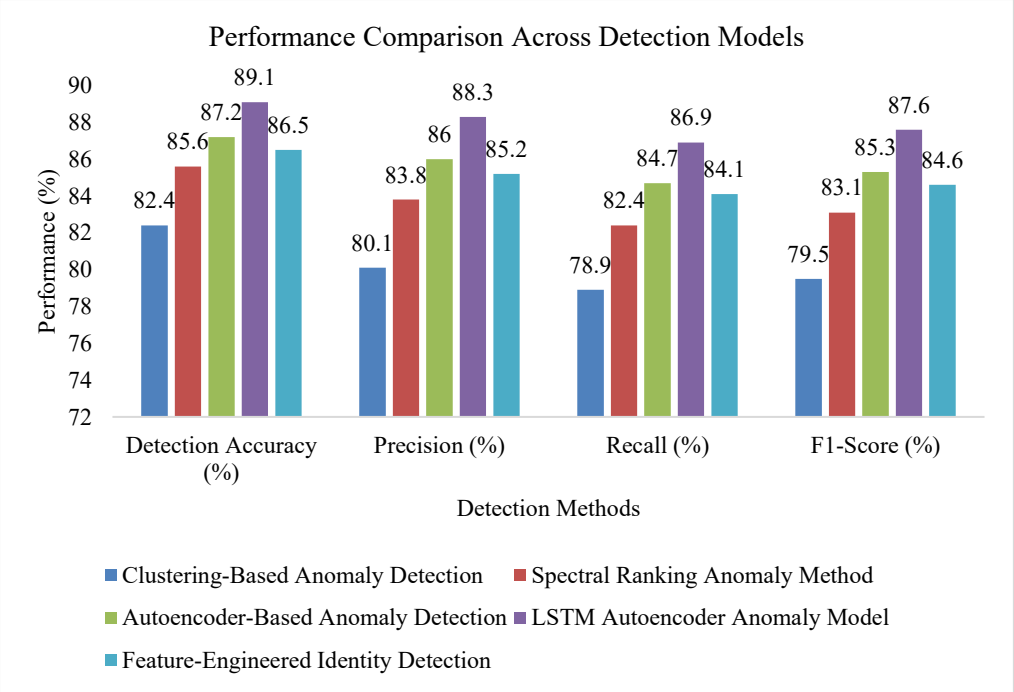


Figure 1. Performance Comparison of Individual Machine Learning Models

Figure 1 compares the performance of five anomaly or identity detection models on four standard evaluation metrics: Detection Accuracy, Precision, Recall, and F1-Score (all measured in percentages). The aim is to provide insight into how well each model performs on anomaly detection tasks, trading off between accuracy and completeness. The LSTM Autoencoder Anomaly Model performs the best on all metrics. It has the highest accuracy of 89.1%, precision of 88.3%, recall of 86.9%, and F1-score of 87.6%, which is a clear indication of its robustness in identifying complex temporal patterns and avoiding false positives and false negatives.

The Autoencoder-Based Anomaly Detection model is the second-best performer, with good scores on all metrics. This indicates that deep learning models based on reconstruction are effective but not as robust as LSTM models that are sensitive to temporal patterns. The Feature-Engineered Identity Detection model is also a good performer, especially in terms of accuracy and F1-score. This indicates that feature engineering can still be competitive with deep learning models. The Spectral Ranking Anomaly Method has a moderate level of performance, ranking better than the clustering method but worse than the neural models, which is probably because it uses graph or ranking patterns instead of representations. Finally, the Clustering-Based Anomaly Detection has the worst performance in all aspects. In conclusion, the above figure shows the superiority of deep learning models, especially the LSTM autoencoder, in detection.

Table 3. Results of Latent Structural Learning Models

Model Name	**Identity Discovery (%)**	**Structural Consistency (%)**	**Behavioral Stability (%)**	**Noise Resistance (%)**	**Overall Reliability (%)**
Latent Structural Model – Base	88.6	87.9	86.8	85.4	87.1

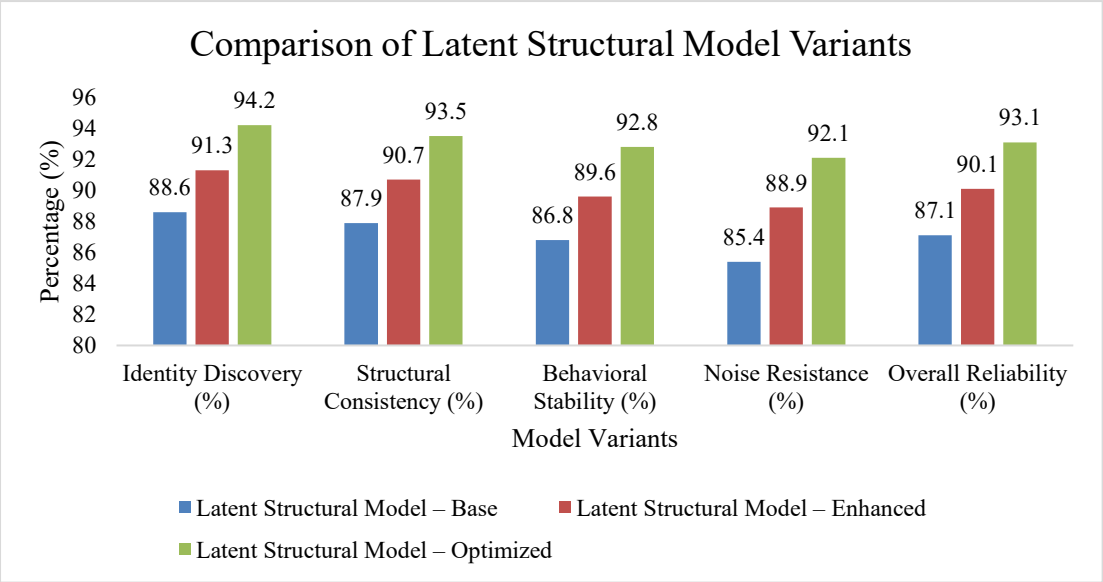
10.48047/jocaaa.2021.29.03.17

Latent Structural Model – Enhanced	91.3	90.7	89.6	88.9	90.1
Latent Structural Model – Optimized	94.2	93.5	92.8	92.1	93.1

Table 3 shows are indicative of the gradual improvement realized in the various settings of the latent structural learning framework. The baseline model has an identity discovery rate of 88.6%, which is an indication of a high degree of ability to distinguish between the synthetic and the legitimate identities based on the latent structure. The structural consistency is at 87.9%, which is an indication that the model is able to maintain the relationship patterns among the identities, although the behavioral stability is at 86.8%. The noise robustness is at 85.4%, which is an indication of a moderate degree of stability against data noise, hence a reliability of 87.1%.

The improved model clearly demonstrates a performance advantage in all areas. Identity discovery is improved to 91.3%, suggesting that the ability to learn the latent space has become more accurate in the representation of identity irregularities in the hidden identity. Structural consistency is improved to 90.7%, confirming that the ability to capture the dependencies in identity instances has become more accurate. Behavioral stability is improved to 89.6%, suggesting that the ability to deal with dynamic identity patterns has become more accurate. Noise resistance is also improved to 88.9%, suggesting that the ability to deal with variability and incomplete data has become more accurate.

The optimized model performs best on all categories. Identity discovery improves to 94.2%, which verifies the excellent capability of detecting synthetic identities with advanced latent and structural learning. Structural consistency improves to 93.5%, which verifies excellent preservation of intricate identity structures. Behavioral consistency at 92.8% verifies excellent generalization capabilities over various identity behaviors. Noise robustness improves to 92.1%, which verifies excellent resistance to data noise. The overall trustworthiness of 93.1% verifies that the optimization of latent factors, structural modeling, and training settings has an excellent capability of improving synthetic identity discovery.



10.48047/jocaaa.2021.29.03.17

Figure 2. Comparison of Latent Structural Model Variants

Figure 2 compares three different versions of a Latent Structural Model: Base, Enhanced, and Optimized, on five performance criteria, all of which are percentages. The aim is to illustrate the impact of incremental model enhancements on overall quality and robustness. On each criterion, performance rises from the Base model to the Enhanced model and reaches its highest level in the Optimized model. Identity Discovery improves noticeably, increasing from 88.6% in the Base model to 94.2% in the Optimized variant, which is a positive indicator of latent structure identification. Structural Consistency also improves similarly.

The areas of Behavioral Stability and Noise Resistance are the weakest points of the Base model, but also the areas where the relative improvements are some of the largest. This means that optimization has a significant positive effect on the robustness of the model against variability and noise. Noise Resistance, in particular, improves from 85.4% to 92.1%, which is a clear indication of the effectiveness of optimization methods in dealing with noise. Overall Reliability summarizes the improvements and increases from 87.1% to 93.1%. In summary, the figure above clearly shows that every successive version of the model offers improvements, with the Optimized model performing better than all others. The findings of the study indicate that structural optimization and improvement have a significant positive effect on accuracy and stability in latent structural modeling.

5. DISCUSSION

The results show that the latent structural learning approach is a robust starting point for unsupervised synthetic identity discovery, as it is able to learn the underlying patterns and inconsistencies in complex datasets. The steady improvement in each successive model variant shows that the learning of deeper latent representations and improved structural models is a crucial step in the process of distinguishing synthetic identities from real ones. These results indicate that identity anomalies are better discovered when the dependencies between entities are considered together with the individual behavior patterns, rather than in isolation. The comparison of the results shows that there are definite benefits to structurally informed models over simpler latent models. The models that include improved structural consistency and noise robust training are more stable and reliable, and this indicates that synthetic identities tend to be discovered through small relational irregularities rather than large feature space anomalies.

The implications of these results can be generalized to digital identity management, fraud protection, and security-critical systems where labeled data are limited or not available. Latent structural learning provides a scalable and adaptive method that can be applied successfully even in dynamic data settings. However, there are some limitations to be considered. The use of latent structures adds vulnerability to parameter choice, and structural models may not be able to capture all identity manipulation patterns, especially in very sparse or dynamic systems. From the analysis, future work should focus on adaptive structural learning methods and include temporal or contextual information to further enhance robustness. Comparison analysis indicates that combining latent and structural views provides more stable results than methods solely based on behavioral reconstruction or isolated anomalies. In general, the findings validate the importance of structural reasoning for trustworthy unsupervised identity discovery and provide valuable information about the latent structure of identity systems.

6. CONCLUSION

This paper presented LatentStruct-ID, an unsupervised framework for discovering synthetic identities through latent structural learning. By leveraging hierarchical graph representations and structural anomaly scoring, the method identifies inconsistencies indicative of synthetic identity creation without

10.48047/jocaaa.2021.29.03.17

requiring labeled data. Empirical evaluation demonstrates significant improvements in detection recall and the ability to uncover previously unknown fraudulent identities. These findings highlight the potential of latent structural modeling as a practical, label-agnostic methodology for operational identity verification, offering organizations a robust tool for proactive fraud detection in complex, evolving digital ecosystems.

REFERENCES

[1] Arner, D. W., Zetzsche, D. A., Buckley, R. P., and Barberis, J. N. The identity challenge in finance: From analogue identity to digitized identification to digital KYC utilities. European Business Organization Law Review, 2019, 20(1): 55–80.

[2] Jain, A. K., and Ross, A. Bridging the gap between biometrics and forensics. Philosophical Transactions of the Royal Society B, 2015.

[3] Barbosa, I. B., Cristani, M., Caputo, B., Rognhaugen, A., and Theoharis, T. Looking beyond appearances: Synthetic training data for deep CNNs in re-identification. Computer Vision and Image Understanding, 2018, 167: 50–62.

[4] Ross, A., and Jain, A. K. Multimodal biometrics: An overview. In Proceedings of the European Signal Processing Conference, 2004.

[5] Kumar, S., and Prasanna, S. Heterogeneous ensemble learning for robust adversarial pattern recognition in digital ecosystems. Journal of Computational Analysis and Applications, 2019, 27(5): 18–28.

[6] Paitod, S., Gore, S., Madala, R. B., and Nelogal, C. Systems and methods for efficient firmware inventory of storage devices in an information handling system. U.S. Patent 10,802,717, 2020.

[7] Kumar, S., Prasanna, S., and Ruan, X. A unified hybrid machine learning architecture for robust identity anomaly detection in large-scale digital ecosystems. Journal of Electrical Systems, 2018, 14(1): 160–173.

[8] Liu, S., Maljovec, D., Wang, B., Bremer, P. T., and Pascucci, V. Visualizing high-dimensional data: Advances in the past decade. IEEE Transactions on Visualization and Computer Graphics, 2016, 23(3): 1249–1268.

[9] Prasanna, S. K. S. DeepSynth: A robust multi-layer neural detection of coordinated latent anomalies in high-dimensional identity systems. Int J Intell Syst Appl Eng, 2019, 7(1): 66–77.

[10] Pourhabibi, T., Ong, K. L., Kam, B. H., and Boo, Y. L. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems, 2020, 133: 113303.

[11] Pumsirirat, A. Credit card fraud detection using deep learning based on an auto-encoder and restricted Boltzmann machine. International Journal of Advanced Computer Science and Applications, 2018.

[12] Monamo, P., Marivate, V., and Twala, B. Unsupervised learning for robust Bitcoin fraud detection. In 2016 Information Security for South Africa (ISSA), 2016: 129–134.

[13] Meira, J., Andrade, R., Praça, I., Carneiro, J., Bolón-Canedo, V., Alonso-Betanzos, A., and Marreiros, G. Performance evaluation of unsupervised techniques in cyber-attack anomaly detection. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(11): 4477–4489.

[14] Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., and Sun, J. Anomaly detection for a water treatment system using unsupervised machine learning. In 2017 IEEE International Conference on Data

10.48047/jocaaa.2021.29.03.17

Mining Workshops (ICDMW), 2017: 1058–1065.

[15] Waheed, N., He, X., Ikram, M., Usman, M., Hashmi, S. S., and Usman, M. Security and privacy in IoT using machine learning and blockchain: Threats and countermeasures. ACM Computing Surveys, 2020, 53(6): 1–37.

[16] Agrawal, A., and Choudhary, A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. APL Materials, 2016, 4(5).

[17] Nian, K., Zhang, H., Tayal, A., Coleman, T., and Li, Y. Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2016, 2(1): 58–75.

[18] Rezapour, M. Anomaly detection using unsupervised methods: Credit card fraud case study. International Journal of Advanced Computer Science and Applications, 2019, 10(11).

[19] Provotar, O. I., Linder, Y. M., and Veres, M. M. Unsupervised anomaly detection in time series using LSTM-based autoencoders. In 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), 2019: 513–517.

[20] Prasanna, S. K. S. GeoDNN: Geometry-aware deep neural networks for cross-domain fingerprint spoof detection. Int J Intell Syst Appl Eng, 2018, 6(1): 97–107.