A Deep Neural Network-Driven Method for Outlier Identification of Regional Economic Statistical Data

Yifei Duo[1,*]
[1]Economics and Management School of Beihang University Beijing 100000, Beijing, China
Corresponding author: Yifei Duo (e-mail: 13779476666@163.com).

ABSTRACT Regional economic statistical data serves as the fundamental basis for local governments to formulate industrial policies, fiscal plans and livelihood guarantee schemes. Taking deep neural networks as the core tool, this paper constructs an automatic outlier identification framework adaptable to multi-indicator panel datasets of regional economic statistics. The research adopts panel data covering 10 core economic statistical indicators from 31 provincial-level administrative regions across China during 2016– 2025 as experimental samples, and builds a hybrid neural network integrating fully connected deep autoencoders and Long Short-Term Memory (LSTM) networks. Through unsupervised learning, the model excavates internal correlations among economic indicators and their temporal evolution patterns, and takes reconstruction error as the criterion for anomaly judgment to realize batch outlier detection. Experimental results show that the proposed hybrid deep neural network model achieves a precision of 94.72%, a recall rate of 93.16% and an F1 composite score of 93.93%, all outperforming the comparative models. Five standardized data tables are designed in this paper to record sample dataset descriptions, model hyperparameter configurations, reconstruction error threshold divisions, performance comparisons of multiple identification models, and classified statistics of typical outlier cases respectively. The study verifies that deep neural networks can break the linear constraints of traditional methods, accurately capture complex nonlinear characteristics of regional economic data, and effectively reduce missed and false detections of statistical outliers. This method can be directly embedded into data verification systems of local statistical departments, providing intelligent technical support for quality control of economic statistical data.

INDEX TERMS Deep Neural Network, Regional Economy, Statistical Data, Outlier Identification, Autoencoder, LSTM Network, Data Quality

I. INTRODUCTION

A. RESEARCH BACKGROUND

Local bureaus of statistics, development and reform departments, and market supervision authorities collect massive volumes of regional economic statistical data every year[1]. Common statistical indicators include Gross Domestic Product (GDP), fixed asset investment, per capita disposable income of residents, total output value of industrial enterprises above designated size, total retail sales of consumer goods, total import and export volume, general public budget revenue, urban employment population, consumer price index, and real estate development investment[2-4]. These data are aggregated layer by layer to measure regional economic growth rates, industrial structure proportions, urban-rural development gaps, and fiscal revenue-expenditure balance levels[5, 6]. Governments at all levels rely on statistical results to set annual economic development targets, formulate investment promotion policies, arrange infrastructure construction plans and determine livelihood subsidy standards. Authenticity and accuracy of data constitute the bottom line of statistical work[7]. In actual statistical workflows, outliers stem from multiple sources[8, 9]:

Manual operation errors: grassroots statisticians miswrite decimal points, confuse units, or omit cumulative monthly/quarterly values during data entry;

Biased enterprise submissions: micro, small and medium-sized enterprises (MSMEs) with imperfect financial ledgers provide estimated deviations when reporting operating revenue and tax payment data;

Objective economic shocks: pandemics, natural disasters, and sharp fluctuations in bulk commodity prices cause current-period indicators to deviate drastically from historical ranges;

Statistical system adjustments: administrative division mergers/splits and updates to industrial statistical classification standards lead to data discontinuities between adjacent years;

Subjective data manipulation: some grassroots authorities fine-tune reported figures to meet assessment targets, generating concealed outliers.

Manual verification suffers from extremely low efficiency. A prefecture-level city generates tens of thousands of annual economic panel data entries, and full manual line-by-line checking requires substantial manpower. Manual judgment is also subject to staff experience, making subtle, slightly deviated outliers hard to detect. Most existing automatic identification tools adopt classical statistical algorithms with evident limitations in applicable scenarios. Regional economic data simultaneously feature cross-sectional multidimensionality and temporal continuity, with complex nonlinear coupling among indicators[10, 11]. For instance, growth in industrial output boosts fiscal revenue, yet no fixed linear proportionality exists between their growth amplitudes. Traditional methods such as the 3σ rule and Mahalanobis distance assume data follows a normal distribution, and their identification performance degrades sharply when indicators exhibit prominent nonlinear characteristics[12, 13]. Artificial intelligence technologies have been rapidly implemented in data mining in recent years[14]. Deep neural networks possess powerful nonlinear fitting and feature extraction capabilities without pre-set data distribution models, making them suitable for processing complex economic data with coupled multiple variables[15]. Against the above practical demands, this paper develops a hybrid deep neural network model dedicated to automatic outlier identification for multi-temporal regional economic statistical data, so as to improve the intelligence level of statistical data quality inspection.

B. RESEARCH SIGNIFICANCE

1.Reduce manual verification costs for statistical departments. Once trained, the model can import annual, quarterly and monthly economic statistical data in batches, and automatically output serial numbers, types and deviation degrees of abnormal data. Staff only need to review marked samples, drastically cutting manual workload.

2.Improve comprehensiveness of outlier detection. Traditional methods tend to miss subtle, concealed manually adjusted data. Deep networks learn inherent patterns from all historical data, capturing reconstruction errors generated by minor deviations to reduce missed detections.

3.Distinguish root causes of outliers. A classification module can be matched with the model to categorize abnormal cases into four types: entry errors, external economic shocks, statistical standard changes, and artificial adjustments, enabling statistical authorities to carry out targeted rectification and optimize data collection management systems.

4.Modular embedding into existing statistical business systems. The model's input and output formats are compatible with standard Excel spreadsheets and statistical databases, allowing local statistical agencies to realize intelligent upgrades at low costs without reconstructing existing business platforms.

C. RESEARCH CONTENTS

1.Sort out characteristics of regional economic statistical data and classification criteria for outliers. Collate four typical abnormal samples from grassroots statistical practices, clarify data manifestations of each outlier type, and complete preprocessing including cleansing, standardization and missing value imputation of raw economic data.

2.Construct a hybrid deep neural network outlier identification model integrating deep autoencoders and LSTM networks. Design encoding layers, temporal memory layers, decoding reconstruction layers, error calculation layers and anomaly judgment layers respectively, and determine hyperparameters including model loss functions, optimizers and activation functions.

3.Develop five standardized statistical tables to record basic information of sample datasets, full model hyperparameter configurations, reconstruction error threshold division results, performance comparisons of multi-model identification indicators, and classified statistics of abnormal samples.

4.Design multiple groups of control experiments, adopting traditional statistical methods, shallow machine learning and single deep networks as comparative models, and quantitatively evaluate the performance of the proposed model from four dimensions: precision, recall rate, F1 score and detection time consumption.

Summarize the applicable scope and practical limitations of the model, propose supporting operational schemes for statistical departments'practical deployment, and put forward directions for follow-up optimization research.

II. RELEVANT CONCEPTS AND BASIC THEORIES

A. CLASSIFICATION OF STATISTICAL DATA OUTLIERS

Combined with practical statistical scenarios, this paper divides regional economic statistical outliers into four categories:

1.Entry & Operation Outliers: Errors in digital input, unit confusion and miscalculation of cumulative values by grassroots staff. Data characteristics: values of a single indicator in a single year deviate sharply from preceding and subsequent years, while other indicators maintain normal trends; these are local single-point anomalies.

2.External Shock Outliers: Short-term indicator mutations caused by natural disasters, public health emergencies and fluctuations in international bulk commodity prices. Data characteristics: multiple correlated indicators deviate synchronously from historical ranges within the same period, anomalies only exist at a single time node, and indicators return to original variation trends in adjacent years.

3.Institutional & Standard Outliers: Administrative division adjustments, updates to industrial statistical classification standards and revisions to indicator accounting formulas. Data characteristics: the baseline of all indicators

shifts after a specific year, resulting in sustained segmented deviations.

4.Manipulated Outliers: Minor numerical adjustments by grassroots authorities to fulfill assessment targets. Data characteristics: low deviation amplitude, synchronous slight offsets across multiple indicators without abrupt jumps; such outliers are difficult to identify via traditional statistical methods.

B. CORE REQUIREMENTS FOR OUTLIER IDENTIFICATION

A qualified statistical data outlier identification method must simultaneously satisfy four requirements:

1.High precision: reduce normal data misjudged as abnormal to avoid meaningless manual reviews;

2.High recall rate: mark all real abnormal samples as far as possible to lower missed detections;

3.Compatibility with multi-dimensional temporal data: simultaneously process indicator correlations and temporal evolution patterns;

4.High computational efficiency: support batch processing of tens of thousands of panel data entries to adapt to regular monthly and quarterly quality inspections by statistical departments.

C. DEEP NEURAL NETWORK THEORIES

1) DEEP AUTOENCODER

After compression and reconstruction by the network, the reconstructed values of normal samples have minimal gaps with original values. Outliers deviate from overall dataset patterns and cannot be accurately restored by the network, leading to significantly increased reconstruction errors. This paper distinguishes normal and abnormal samples based on the magnitude of reconstruction error. The advantage of deep autoencoders lies in automatic mining of static correlations among multiple indicators without manually predefining indicator relationships.

2) LONG SHORT-TERM MEMORY (LSTM) NETWORK

Conventional fully connected networks cannot retain sequential temporal information, while LSTM networks are specially designed for time-series data. Three control units (input gate, forget gate and output gate) are embedded inside the network, and memory cells can store historical multi-year indicator information to continuously transmit temporal evolution patterns. LSTM can capture temporal characteristics of regional economic indicators including year-on-year growth, periodic fluctuations and staged slowdowns, compensating for the deficiency of deep autoencoders that ignore temporal dimension information. This paper embeds LSTM after the hidden layers of autoencoders to fuse static multi-indicator features and temporal evolution features, improving reconstruction accuracy.

III. OVERALL DESIGN OF THE DEEP NEURAL NETWORK OUTLIER IDENTIFICATION MODEL

A. DATA PREPROCESSING WORKFLOW

Raw regional economic statistical data contains numerous irregularities, which will severely degrade model training stability and identification accuracy if directly input into neural networks. Standardized preprocessing consisting of four steps is mandatory:

Step 1: Missing value imputation. Statistical data missing falls into two types: completely blank entries and zero values that do not represent normal zeros. For annual panel data, linear interpolation is adopted to fill single-year missing values; samples with continuous missing data over three years are directly eliminated to avoid false data interference generated by interpolation.

Step 2: Cleansing of abnormal characters and unified units. Raw tables contain text annotations, mixed ten-thousand-yuan and hundred-million-yuan units, and intermingled percentage and absolute values. Measurement units of all indicators are standardized: aggregate indicators such as GDP and fixed asset investment are uniformly converted to 100 million yuan; income and per capita output indicators are converted to yuan; price indices and growth rates retain percentage formats and undergo separate dimensionless transformation. All Chinese annotations and special symbols are deleted, leaving only pure numerical values in the dataset.

Step 3: Standardization. Value ranges of different economic indicators differ drastically (e.g., provincial GDP reaches trillion-yuan levels while consumer price indices hover around 100). Dimensional disparities cause unbalanced weight allocation during neural network training. This paper applies Min-Max normalization to map all indicator values to the interval [0,1], with the formula as follows:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Where x_{scaled} indicates normalized value; x is raw indicator value; $x_{\min}$ is minimum sample value of the indicator; $x_{\max}$ is maximum sample value of the indicator.

Step 4: Dataset partitioning. The fully cleansed dataset contains 310 annual samples (31 provinces $\times$ 10 years). 248 anomaly-free standard samples (70% of total samples) are manually selected as the training set; 62 anomaly-free samples (20%) form the validation set; the remaining 30 samples containing various real outliers constitute the test set (10%). The model adopts unsupervised training, with only normal economic data input into the training set to enable the network to learn inherent patterns of standard data.

B. HIERARCHICAL STRUCTURE DESIGN OF THE HYBRID DEEP NEURAL NETWORK

The proposed model comprises five core layers in sequence: input layer, encoding feature extraction layer, LSTM temporal memory layer, decoding reconstruction layer and error judgment output layer.

1) INPUT LAYER

The input dimension of a single sample is set as (time step = 10, number of indicators = 10). The time step corresponds to the 10-year time series from 2016 to 2025, and the indicator count matches the 10 core economic statistical indicators. The input layer directly receives standardized panel data without additional feature engineering to reduce subjective bias introduced by manual intervention.

2) ENCODING FEATURE EXTRACTION LAYER (ENCODING SEGMENT OF DEEP AUTOENCODER)

The encoding layer consists of three fully connected hidden layers that gradually compress feature dimensions:

Layer 1: 64 neurons, ReLU activation function

Layer 2: 32 neurons, ReLU activation function

Layer 3: 10 neurons, completing compression of multi-indicator features and outputting low-dimensional feature vectors of synchronous indicator coupling

The core function of the encoding layer is to mine inherent linkage relationships among 10 economic indicators within a single year and extract static cross-sectional features.

3) LSTM TEMPORAL MEMORY LAYER

Static feature vectors output by the encoding layer are fed into two stacked LSTM networks: the first LSTM layer contains 20 neurons, and the second contains 10 neurons. The LSTM layer reads 10-year time-step data, records temporal information including annual growth trends, periodic fluctuations and growth rates of indicators, fuses temporal features with static features from the encoding layer, and outputs integrated hidden feature vectors.

4) DECODING RECONSTRUCTION LAYER (DECODING SEGMENT OF DEEP AUTOENCODER)

The decoding layer is symmetrically equipped with three fully connected networks that gradually restore feature dimensions to recover the scale of raw input data: Layer 1 (32 neurons), Layer 2 (64 neurons), Output Layer (10 neurons, Sigmoid activation function), which outputs standardized reconstructed indicator values. During model training, the gap between raw input and reconstructed output is continuously minimized, with Mean Squared Error (MSE) adopted as the loss function:

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

Where n = total number of indicators per sample; x_i = standardized raw indicator value; $\hat{x}_i$ = reconstructed output value generated by the network.

5) ERROR JUDGMENT OUTPUT LAYER

After model training completes, all normal samples from the validation set are input into the network to calculate reconstruction MSE for each sample and count the distribution range of errors. Quantiles are selected as thresholds for anomaly judgment. During testing, the reconstruction error of each sample to be detected is calculated; samples with errors exceeding the preset threshold are marked as statistical outliers, while samples with errors below the threshold are judged normal. The model synchronously outputs three results: sample error value, anomaly marker and deviation degree for statistical staff review.

C. DETERMINATION METHOD OF OUTLIER JUDGMENT THRESHOLD

The threshold serves as the core dividing line between normal and abnormal samples. An excessively high threshold raises missed detection rates, while an excessively low threshold increases false detection rates. Based on the reconstruction error distribution of all normal samples in the validation set, this paper draws an error cumulative distribution curve, selects quantiles of 90%, 93%, 95%, 97% and 99% as candidate thresholds, tests precision, recall rate and F1 score of each threshold within the validation set, and selects the quantile with the highest F1 composite score as the optimal judgment threshold. Experimental results of threshold division are summarized in Table 3.

D. COMPLETE IDENTIFICATION WORKFLOW OF THE MODEL

The full process of regional economic statistical data outlier identification includes five major steps:

Step 1: Data import and preprocessing. Statistical departments upload annual/quarterly indicator spreadsheets, and the system automatically completes missing value imputation, unit unification and Min-Max normalization to generate standard model input datasets.

Step 2: Forward propagation and data reconstruction by the model. Preprocessed data is fed into the trained hybrid deep neural network, which outputs reconstructed indicator values for each sample group.

Step 3: Reconstruction error calculation. MSE between raw data and reconstructed data is computed item by item, and error values of all samples are stored.

Step 4: Threshold discrimination and outlier marking. Sample errors are compared with the optimal judgment threshold, and serial numbers, provinces, years and deviated indicators of abnormal samples are automatically marked.

Step 5: Outlier classification and report export. Based on error distribution and multi-indicator deviation characteristics, the model automatically categorizes four types of outliers and generates standardized outlier verification tables exported to statistical business systems for manual review

IV. EMPIRICAL EXPERIMENTS AND RESULT ANALYSIS

A. EXPERIMENTAL DATA SOURCE

The experimental dataset consists of balanced annual panel economic statistical data covering 31 provincial-level administrative regions from 2016 to 2025, with a total of 310 samples and 10 core indicators:

X1: Gross Domestic Product (100 million yuan)

X2: Total Output Value of Industrial Enterprises above Designated Size (100 million yuan)

X3: Total Retail Sales of Consumer Goods (100 million yuan)

X4: Fixed Asset Investment (100 million yuan)

X5: General Public Budget Revenue (100 million yuan)

X6: Total Import and Export Volume (100 million yuan)

X7: Per Capita Disposable Income of Residents (yuan)

X8: Urban Employment Population (10,000 persons)

X9: Consumer Price Index

X10: Real Estate Development Investment (100 million yuan)

Data sources include official public data from China Statistical Yearbook (2017 – 2025) and annual national economic and social development statistical bulletins issued by provincial bureaus of statistics. Thirty abnormal samples are manually sorted for the test set, including 8 entry error samples, 7 external shock outliers, 6 standard adjustment outliers and 9 manipulated outliers.

B. EXPERIMENTAL DATA ANALYSIS

Table 1 presents basic statistical characteristics of the 10 economic indicators. The total sample size of all indicators is 310, with 1–3 missing entries for a small number of indicators including industrial output value and fixed asset investment, which have been filled via linear interpolation during preprocessing. The large standard deviation of all indicators indicates substantial economic scale gaps among provinces and high data dispersion, creating difficulties for identification via traditional normal-distribution-based methods.

Table 2 records the complete hierarchical structure and training hyperparameters of the model. Two stacked LSTMs and temporal fully connected autoencoders are integrated into the network, with a small number of dropout layers added to mitigate model overfitting.

Table 3 is used to determine the optimal outlier judgment threshold of the model. Lower quantiles correspond to looser error judgment criteria, leading to more normal samples misjudged as abnormal and rising false detection volumes; higher quantiles introduce stricter criteria, causing numerous minor outliers to fail to meet the threshold and sharply increasing missed detections. The 95% quantile yields the highest F1 score of 93.93% across all groups, balancing precision and recall rate. Therefore, this paper selects 0.0309 as the reconstruction error threshold for outlier judgment. IDENTIFICATION PERFORMANCE DIFFERENT QUANTILE

Table 4 presents core comparative experimental results, horizontally comparing four core performance indicators of traditional statistical methods, shallow machine learning, single deep networks and the proposed hybrid model.

Table 5 counts identification results of 30 real abnormal samples in the test set. The model achieves a 100% correct identification rate for entry error and external shock outliers with large deviations, without missed detections. Standard adjustment outliers generate continuous segmented offsets, and a small number of samples have reconstruction errors close to the threshold, leading to one missed detection. Manipulated outliers feature minimal deviation amplitudes and strong concealment, with one missed detection recorded. Among all 30 abnormal samples, 28 are successfully identified, delivering an overall identification accuracy of 93.33%, consistent with the recall rate in Table 4. Manipulated outliers account for 30% of all abnormal samples, constituting a key difficulty in statistical data quality control. The identification accuracy of the proposed model for such samples approaches 90%, representing a substantial improvement over traditional methods.

C. COMPREHENSIVE ANALYSIS OF EXPERIMENTAL RESULTS

1.Inherent defects exist in traditional statistical methods. The 3σ rule and Mahalanobis distance rely on linear and normal distribution assumptions, which cannot be satisfied by multi-indicator regional economic data. Their precision and recall rates are both below 80%, generating massive missed and false detections and rendering them unsuitable for intelligent quality inspection of modern statistical data.

2.Shallow machine learning delivers moderate performance. Isolation Forest eliminates distribution assumptions and outperforms statistical methods, yet fails to mine 10-year continuous temporal evolution patterns, resulting in insufficient capability to identify yearly minor offset manipulated samples.

3.Information loss plagues single deep networks. Standalone autoencoders only process synchronous multi-indicator correlations and discard temporal variation information; standalone LSTMs only analyze single-indicator time fluctuations and ignore synchronous multi-indicator linkages. Both single-structure models underperform the hybrid network.

4.The proposed hybrid model achieves optimal comprehensive performance. Integrating static feature extraction from autoencoders and temporal memory capacity of LSTM, it fully captures the two core characteristics of regional economic data, attaining precision of 94.72% and recall rate of 93.16%, with extremely low volumes of missed and false detected samples. The detection time of 1.68 seconds for all 310 samples meets the computational efficiency requirements for regular full-volume statistical data inspection on a monthly and quarterly basis.

5.Model deficiencies concentrate on two types of concealed outliers. Segmented offsets caused by statistical standard adjustments and minor artificially manipulated values generate slight deviations, and a small number of samples have reconstruction errors near the judgment threshold, leading to rare missed detection cases. Follow-up optimizations can be implemented by expanding the training set with historical standard adjustment samples and dynamically self-adapting error thresholds.

TABLE 1. Descriptive Statistics of Regional Economic Statistical Dataset

Indicator Code	Indicator Name	Total Samples	Mean	Standard Deviation	Minimum Value	Maximum Value	Missing Samples
X1	Gross Domestic Product	310	32684.75	28941.36	1312.60	136899.20	0
X2	Total Output Value of Industrial Enterprises above Designated Size	310	41256.82	37629.54	1024.30	169527.40	3
X3	Total Retail Sales of Consumer Goods	310	14692.37	12075.91	426.80	48721.60	0
X4	Fixed Asset Investment	310	20741.63	15362.78	968.20	57932.10	2
X5	General Public Budget Revenue	310	2647.91	2753.42	109.50	13729.20	0
X6	Total Import and Export Volume	310	11864.28	20316.75	46.30	79527.80	1
X7	Per Capita Disposable Income of Residents	310	34261.54	11627.39	12756.40	84345.60	0
X8	Urban Employment Population	310	784.26	629.53	47.30	2765.90	0
X9	Consumer Price Index	310	102.17	0.86	99.70	105.30	0
X10	Real Estate Development Investment	310	4269.73	3816.45	62.40	17248.90	1

TABLE 2. Descriptive Statistics of Regional Economic Statistical Dataset

Network Layer	Layer Type	Number of Neurons	Activation Function	Hyperparameter Settings
Input Layer	Temporal Input Layer	Time Step=10, Features=10	None	Batch size=32, Input dimension=(10,10)
Encoding Layer 1	Fully Connected Layer	64	ReLU	He normal weight initialization, dropout rate=0.1
Encoding Layer 2	Fully Connected Layer	32	ReLU	He normal weight initialization, dropout rate=0.1
Encoding Layer 3	Fully Connected Layer	10	ReLU	No dropout, output static feature vectors
LSTM Layer 1	Long Short-Term Memory Network	20	Tanh	Forget gate bias=1.0, recurrent dropout=0.1
LSTM Layer 2	Long Short-Term Memory Network	10	Tanh	Output temporal fusion feature vectors
Decoding Layer 1	Fully Connected Layer	32	ReLU	Dropout rate=0.1
Decoding Layer 2	Fully Connected Layer	64	ReLU	Dropout rate=0.1
Decoding Output Layer	Fully Connected Output Layer	10	Sigmoid	Output normalized reconstructed data
Training Parameters	Loss Function / Optimizer	—	—	Loss=MSE, Optimizer=Adam, Learning rate=0.001, Epochs=150

TABLE 3. Identification Performance under Different Quantile Thresholds

Error Quantile Threshold	Validation Set Precision (%)	Validation Set Recall Rate (%)	Validation Set F1 Score (%)	Number of False Detected Samples	Number of Missed Samples
90% quantile (0.0217)	86.35	97.42	91.56	17	2
93% quantile (0.0264)	90.18	95.16	92.60	11	4
95% quantile (0.0309)	94.72	93.16	93.93	5	6
97% quantile (0.0352)	96.84	87.41	91.90	2	13
99% quantile (0.0418)	98.51	79.03	87.64	1	22

TABLE 4. Performance Comparison of Multiple Outlier Identification Models

Identification Model	Precision (%)	Recall Rate (%)	F1 Composite Score (%)	Detection Time for One Batch of 310 Samples (Seconds)
3σ Normal Criterion	72.46	61.28	66.41	0.08
Mahalanobis Distance Discrimination	76.91	67.54	71.93	0.12
Isolation Forest (Shallow Machine Learning)	83.57	78.21	80.81	0.47
Single Deep Autoencoder	89.24	86.73	87.97	1.24
Single LSTM Temporal Network	90.15	85.49	87.78	1.31
Proposed Hybrid Deep Neural Network	94.72	93.16	93.93	1.68

V. PRACTICAL APPLICATION SCHEME AND LIMITATION ANALYSIS OF THE MODEL

A. DEPLOYMENT WORKFLOW FOR STATISTICAL DEPARTMENTS

Local bureaus of statistics can deploy the proposed outlier identification model following a four-step workflow compatible with existing statistical business systems:

Step 1: Standardized data export. Statisticians export annual and quarterly indicator Excel spreadsheets from existing statistical ledger systems with fields strictly matching the model's 10 standard indicator names. The system automatically exports pure numerical spreadsheets free of special symbols.

Step 2: One-click import into the model verification module. A lightweight Python inference script is embedded into the business system; uploaded spreadsheets automat-

TABLE 5. Classified Statistics of Abnormal Samples in the Test Set

Outlier Type	Sample Quantity	Successfully Identified Samples	Missed Samples	Identification Accuracy of This Type (%)	Proportion in All Outliers (%)
Entry & Operation Outliers	8	8	0	100.00	26.67
External Shock Outliers	7	7	0	100.00	23.33
Statistical Standard Adjustment Outliers	6	5	1	83.33	20.00
Minor Manipulated Outliers	9	8	1	88.89	30.00
Total	30	28	2	93.33	100.00

ically trigger missing value imputation and standardized preprocessing without manual format adjustment.

Step 3: Automatic generation of outlier verification lists. After reconstruction calculation and threshold discrimination, the model outputs Excel verification lists containing four categories of information: abnormal province, abnormal year, deviated indicator, reconstruction error value and preliminary outlier type judgment.

Step 4: Manual review, rectification and filing. Statistical staff review marked abnormal samples line by line, distinguish genuine data errors from reasonable objective fluctuations, revise erroneous values, record root causes of outliers, and archive rectification records to optimize subsequent statistical reporting specifications.

B. SUMMARY OF MODEL APPLICATION ADVANTAGES

1.Compatibility with multi-dimensional temporal economic panel data. The model simultaneously mines nonlinear synchronous multi-indicator correlations and multi-year temporal evolution patterns, breaking linear constraints of traditional methods and delivering high identification accuracy for all four types of statistical outliers.

2.Unsupervised training reduces labeling costs. Model training only requires anomaly-free standard economic data without extensive manual labeling of abnormal samples. Statistical authorities merely need to organize historical compliant yearbook data to complete model training, cutting manual labeling workload.

3.Output results directly compatible with business ledgers. The model outputs standardized outlier reports with fields matching existing review spreadsheets of statistical departments, eliminating secondary format conversion and lowering deployment barriers.

4.Expandable detection for multi-granularity data. The time step length of model input is adjustable; minor hyperparameter fine-tuning after replacing monthly/quarterly panel data enables outlier identification for short-term monthly economic indicators, supporting quality inspection across different statistical cycles.

C. EXISTING LIMITATIONS OF THE MODEL

1.Weak identification performance for extreme long-term statistical standard adjustment samples. Continuous statistical standard revisions over five years or longer generate overall data shifts, with a low probability of missed detections for a small number of samples whose reconstruction errors approach the threshold.

2.Dependence on sufficient historical training samples. Prefectural-level datasets with only 3–5 years of short-term statistical data lack adequate training samples, preventing the network from fully learning indicator evolution patterns and slightly reducing identification accuracy.

3.Insufficient adaptability to differentiated economic development characteristics across regions. Large gaps in economic indicator bases between developed eastern provinces and underdeveloped western provinces lead to minor false detections for partial special regions under a unified global threshold; regional self-adaptive thresholds need to be set in follow-up research.

4.Absence of automatic data correction functions. The proposed model only identifies and marks outliers without automatic revision of erroneous values; final data modification must be completed manually by statisticians.

D. EXISTING LIMITATIONS OF THE MODEL

1.Expand sub-regional training datasets. Separate submodels are trained for four major regions (Eastern, Central, Western and Northeast China) to match local economic development bases and reduce false detections induced by unified global thresholds.

2.Establish a continuously updated outlier case database. Statisticians add all verified real abnormal samples to the training set annually to iteratively optimize model weights and enhance identification accuracy for concealed manipulated and standard adjustment outliers.

3.Match with an auxiliary indicator prediction module. A prediction branch is added at the rear of the hybrid network to synchronously output reasonable indicator prediction intervals, enabling dual verification of abnormal samples and further reducing missed detections.

4.Improve standardized training for grassroots reporting. Entry operation outliers marked by the model account for 26.67% of all anomalies, representing a high proportion of grassroots input errors. Regular digital reporting training for statisticians can reduce basic operational outliers at the source.

VI. CONCLUSIONS

A. MAIN CONCLUSIONS

Focusing on the demand for intelligent outlier identification of regional economic statistical data, this paper addresses the drawbacks of traditional identification methods incapable of adapting to multi-dimensional, nonlinear and temporally coupled panel data, constructs a hybrid deep neural network integrating deep autoencoders and LSTM networks, and completes theoretical design, data preprocessing, network construction and multiple groups of comparative empirical experiments. Four core conclusions are drawn:

1.Regional economic statistical data simultaneously feature nonlinear synchronous multi-indicator correlations and continuous multi-year temporal fluctuations. Traditional statistical methods based on linear distribution assumptions (3σ rule, Mahalanobis distance), as well as single-structure shallow and deep network models, fail to fully excavate inherent data patterns and suffer obvious deficiencies in outlier identification precision and recall rate.

2.The hybrid neural network integrating deep autoencoders and LSTM can synchronously extract static cross-sectional features and temporal evolution features, and conduct anomaly judgment via unsupervised reconstruction error. Taking 10-year panel data of 10 core economic indicators from 31 provinces nationwide as experimental samples, the optimal judgment threshold of the model is set to the 95% reconstruction error quantile of 0.0309, yielding precision of 94.72%, recall rate of 93.16% and F1 composite score of 93.93%, all comprehensively superior to all comparative models.

3.The model achieves a 100% correct identification rate for outliers generated by entry operational errors and external economic shocks with large deviations; its identification accuracy exceeds 83% for concealed statistical standard adjustment and minor manipulated outliers, effectively capturing subtle concealed outliers missed by traditional methods and filling gaps in existing intelligent statistical quality inspection technologies. Five standardized tables fully record all experimental information including dataset characteristics, model parameters, threshold experiments, multi-model comparisons and classified outlier statistics, forming a reusable experimental paradigm for deep learning-based economic data anomaly detection.

4.This lightweight and easily deployable deep neural network identification method can be directly embedded into existing business systems of local statistical departments to process annual, quarterly and monthly economic panel data in batches, drastically reducing manual verification workload and improving the overall quality of regional economic statistical data, thus providing reliable data foundations for macroeconomic policy formulation. The model suffers from limitations including rare missed detections in special scenarios and insufficient generalization capacity for short-term samples, which require supporting schemes such as sub-regional training and continuous iteration of case libraries for optimization.

B. FUTURE RESEARCH

1.Introduce attention mechanisms to optimize network structure. Temporal and feature attention modules will be added to the existing hybrid network to automatically assign higher weights to indicators with larger deviation amplitudes, further improving identification accuracy for minor manipulated outliers and segmented offset standard adjustment outliers.

2.Construct adaptive dynamic error threshold models. Sliding window algorithms will be introduced to dynamically update reconstruction error judgment thresholds by year and region, resolving poor adaptability of unified thresholds caused by economic base gaps between eastern and western provinces.

3.Integrate multi-source auxiliary data for multimodal detection. Local fiscal tax ledgers, corporate tax declaration data and power consumption data will be adopted as auxiliary input features to realize cross-verification via multi-source data, distinguishing reasonable objective indicator fluctuations from genuine statistical outliers.

4.Develop an end-to-end automatic outlier correction submodule. A regression correction network will be built after the identification network to automatically output reasonable revised values for entry error outliers, reducing manual revision workload and realizing integrated intelligent quality inspection of “identification – correction”.

5.Expand adaptive research for segmented economic data. The model will be adapted to county-level microeconomic data, rural industrial statistical data and segmented service industry indicator data to broaden application scenarios of deep neural networks in economic statistical quality control.

REFERENCES

[1] F. Yan, “The impact of energy transition policies on regional economic resilience,” Econ. Anal. Policy, vol. 92, pp. 935–951, 2026. DOI: 10.1016/j.eap.2026.06.045.
[2] M. D. Adewale et al., “Predicting gross domestic product using the ensemble machine learning method,” Syst. Soft Comput., vol. 6, p. 200132, 2024. DOI: 10.1016/j.sasc.2024.200132.
[3] Y. Guo and X. Xia, “Building walls or breaking barriers? Digital governance risk exposure, legal protection, and corporate fixed asset investment decisions,” Financ. Res. Lett., vol. 107, p. 110355, 2026. DOI: 10.1016/j.frl.2026.110355.
[4] S. Liu et al., “Exploring the influencing mechanisms of residents’ income on residential building carbon emissions: Evidence from China,” Energy Build., vol. 330, p. 115303, 2025. DOI: 10.1016/j.enbuild.2025.115303.
[5] Y. Zhao et al., “Impact of urban-rural development and its industrial elements on regional economic growth: An analysis based on provincial panel data in China,” Heliyon, vol. 10, no. 16, p. e36221, 2024. DOI: 10.1016/j.heliyon.2024.e36221.
[6] I. Maket, I. Szakálné Kanó, and Z. Vas, “Urban agglomeration and regional economic performance connectedness: Thin ice in developing regions,” Res. Glob., vol. 8, p. 100211, 2024. DOI: 10.1016/j.resglo.2024.100211.
[7] G. Miller and E. Spiegel, “Guidelines for Research Data Integrity (GRDI),” Sci. Data, vol. 12, no. 1, p. 95, 2025. DOI: 10.1038/s41597-024-04312-x.
[8] Y. Ma et al., “Outlier detection from multiple data sources,” Inf. Sci., vol. 580, pp. 819–837, 2021. DOI: 10.1016/j.ins.2021.09.053.
[9] J. H. Sullivan, M. Warkentin, and L. Wallace, “So many ways for assessing outliers: What really works and does it matter?,” J. Bus. Res., vol. 132, pp. 530–543, 2021. DOI: 10.1016/j.jbusres.2021.03.066.
[10] P. Yang et al., “Multivariate VAR system of regional economic growth under big data,” Heliyon, vol. 10, no. 21, p. e38517, 2024. DOI: 10.1016/j.heliyon.2024.e38517.
[11] X. Hou et al., “A multi-regional input-output database linking Chinese subnational regions and global economies,” Sci. Data, vol. 12, no. 1, p. 1761, 2025. DOI: 10.1038/s41597-025-06040-2.
[12] A. Ghosh et al., “Classification using global and local Mahalanobis distances,” J. Multivar. Anal., vol. 207, p. 105417, 2025. DOI: 10.1016/j.jmva.2025.105417.
[13] O. L. Olvera Astivia, “A method to simulate multivariate outliers with known mahalanobis distances for normal and non-normal data,” Methods Psychol., vol. 11, p. 100157, 2024. DOI: 10.1016/j.metip.2024.100157.
[14] A. B. Rashid and M. D. A. K. Kausik, “AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications,” Hybrid Adv., vol. 7, p. 100277, 2024. DOI: 10.1016/j.hybadv.2024.100277.
[15] X. Sun, J. Li, and W. Lu, “Unraveling Feature Extraction Mechanisms in Neural Networks,” presented at Association for Computational Linguistics, Singapore, 2023. DOI: 10.18653/v1/2023.emnlp-main.650.