Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Design of Coal Mine Gas Explosion Early Warning System Based on Big Data

[1]Lei Xu, [2]Siti Norbaya Daud, [3]Fengling Liu
[1]*Faculty of Engineering, Built Environment & Information Technology, SEGi University*
[2]*Southwest Guizhou Vocational &Technical College Nationalities*
Email：18684133266@163.com
[2]*Faculty of Engineering, Built Environment & Information Technology, SEGi University*
Email：norbayadaud@segi.edu.my
[3]*University of Technology MARA(UiTM)*
Email：leen0707@163.com

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

In the continuous updates of modern early warning technologies, traditional coal mine gas explosion early warning methods struggle to ensure data authenticity and periodicity, failing to provide real-time analysis and accurate predictions. In response to the diverse and heterogeneous characteristics of data sources in coal mine safety systems, this study combines big data technology to research and design a coal mine gas explosion early warning system. This system integrates data mining techniques, mathematical derivations, coal mine safety management, and big data analysis methods, featuring data collection, data processing, data storage, risk prediction, visual presentation, and email alerts. The system model and functional framework are provided, with detailed descriptions of the big data implementation processes for each functional module, effectively enhancing the risk prevention capabilities and decision-making efficiency of coal mine gas explosions.

Keywords: Data mining ; Data acquisition ;Data processing;Data storage;Risk profile ;Early warning system

1. INTRODUCTION

Coal mine extraction activities are constrained by the complex and ever-changing production environment, and coal mine disasters pose significant threats to production safety. Among these, coal mine gas explosions, due to their sudden onset and high destructiveness, severely endanger the lives of coal miners and hinder the development of coal mining enterprises. Over the past few decades, coal mining companies have continued to use traditional methods and early warning systems as primary means for preventing and handling coal mine gas explosions. However, due to technological limitations, traditional data collection methods and analysis models mainly rely on manual inspections, sensor-based monitoring, and historical event analysis. Although these methods have been successful in identifying potential risks, they fail to achieve effective capture and handling. For the large volume of multi-source, heterogeneous data generated by coal mine production, traditional methods struggle to ensure data authenticity and periodicity(Li et al,2022;Li,2022;Dong,2019). Therefore, in most cases, they cannot provide real-time analysis or accurate predictions, leading to reactive rather than proactive prevention.

As coal mine production technology continues to advance, the requirements for safe production are constantly increasing. In the future, the interpretation of safety will place greater emphasis on predictive prevention before accidents rather than post-accident investigation and research. Given the diverse and heterogeneous characteristics of data sources in coal mine safety systems, real-time processing and analysis of massive datasets through big data distributed data mining capabilities and robust data analysis models can uncover information and patterns from complex and intricate data, revealing anomalies(Gao et al,2022;Guo,2019;He,2013). This enables multi-dimensional and high-precision prediction of coal mine gas explosion disasters.

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

However, in building a coal mine gas explosion early warning system based on big data, there are still issues that need to be addressed, such as the system's scalability, operability, and the quality of source data. Addressing these issues requires interdisciplinary research collaboration, including expertise from fields like data science, mining engineering, and risk management. In light of this, this paper proposes a big data coal mine gas explosion early warning system, relying on big data platform technology, aiming to analyze coal mine production data and provide technical support for decision-making in coal mine gas explosion early warning.

2. KEY TECHNICAL ISSUES

The establishment of coal mine gas explosion early warning system is a complex system engineering involving software and hardware resource management, optimal configuration and so on. To establish this system, at least the following problems need to be solved.

2.1 Ensure the accuracy of data acquisition

The data generated by coal mine production is mainly composed of basic data and external data. At present, the data obtained by traditional reporting methods have poor authenticity, timeliness and periodicity, so how to enhance the reliability of data is a key factor in data acquisition(Jane,2019;Liu et al,2021).

2.2 Effective processing of data

Data collected by underground sensors is prone to interference from environmental factors, leading to some data being abnormal or missing, which significantly impacts the accuracy of data analysis results. However, discarding these abnormal data points directly would also reduce the sample size, thereby decreasing the precision of data analysis(Cheng & Yu,2021) Therefore, based on existing hardware and software resources, researching and establishing a data processing model using virtualization technology to clean, integrate, standardize, and transform the data, achieving reasonable organization of the data, is a critical prerequisite for scientific data analysis.

2.3 Effective storage of data

Coal mine production data is characterized by large data volumes, diverse data types, and extensive data extension ranges. If traditional relational databases are used for storage, it would lead to issues with the inconvenience of adjusting data structures(Li,2022). Therefore, when selecting a storage method for this data, one should opt for a highly scalable and reliable data storage framework to facilitate future adjustments to the data storage structure.

3. KEY TECHNOLOGIES OF THE EARLY WARNING PLATFORM

The data collection, mining and analysis system built by Hadoop for the risk prediction of coal seam and gas outburst needs to cover many steps from data collection to real-time analysis, historical data comparison and risk prediction.

Using big data preprocessing techniques to clean, integrate, standardize, and transform the data collected by sensors, leveraging the Hadoop computing framework and distributed frameworks to train analytical models on historical data. The trained model is then used to analyze real-time data and generate predictive results. Meanwhile, the original data and analyzed data are stored using a distributed storage framework. Finally, based on the predictive results, alerts and visualizations are generated.

3.1 Hadoop framework

Hadoop is an open source distributed computing framework, whose core architecture is mainly composed of Kafka, HDFS (Hadoop Distributed File System), Spark and MapReduce. It is suitable for big data storage, batch processing, real-time computing, machine learning and other scenarios(Pan et al,2021;Pei et al,2012). The structure of Hadoop project is shown in Figure 1.

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Hadoop Ecosystem		
HDFS (Storage layer)	YARN (Resource Management)	MapReduce (Computational framework)
HBase (NoSQL database)	Hive (Data Warehouse)	Spark (Computing engine)
Pig (Scripting language)	Sqoop (Data migration tool)	Kafka (Data collection)
ZooKeeper (Distributed coordination)	Oozie (Workflow scheduling)	Mahout (Machine learning library)

Figure 1. structure of Hadoop project

The big data cloud platform is a supercomputer based on virtualization technology and carried by the network. It integrates a series of large-scale, scalable computing services, storage data applications, and other resources through basic architecture platforms or integrated tools to work collaboratively. From a technical perspective, its architecture consists of hardware resources, virtual resources, management middleware, and related service interfaces.

3.2 Data acquisition technology

Apache Kafka is a distributed stream processing platform primarily used for building real-time data pipelines and streaming applications. It boasts high throughput, capable of handling millions of messages per second, and can persist messages to disk, supporting data replay for high durability. Performance can be enhanced by adding Broker and Partition, offering excellent scalability. Therefore, it is widely applied in scenarios such as log collection, messaging systems, event tracing, and stream processing(Qiao et al,2020).

In the Kafka component, the Pub-Sub model (publish-subscribe) is adopted. Producers publish messages (such as logs, transaction data, sensor data, etc.) to topics, which are partitioned into multiple partitions to enhance parallelism. Consumers in consumer groups subscribe to these topics and process the messages (such as real-time analysis, data storage, etc.). Messages in each partition are consumed with offset tracking to monitor progress. The replica mechanism Replication prevents data loss, ensuring high availability, while synchronous replication (ISR) ensures data consistency.

3.3 Data processing technology

Apache Spark is a fast and versatile cluster computing system that provides efficient data processing capabilities through elastic distributed datasets (RDD). Its advantages include high-speed in-memory computation, excellent language support (Python, Scala, Java, R, SQL), a powerful unified engine (integrated batch processing, stream processing, machine learning, graph computing), and high fault tolerance for automatic data recovery based on RDD(Wang et al,2021). Spark applications consist of drivers (Driver Program) and executors (Executors), with resource scheduling managed by the cluster manager (Cluster Manager). Spark SQL supports structured data processing, while DataFrame and Dataset API offer more advanced abstractions. Spark Streaming can handle real-time data streams, and MLlib provides machine learning capabilities.

Based on the above introduction, Apache Spark is a high-speed, versatile, and user-friendly big data computing engine that supports batch processing, stream processing, machine learning, graph computation, and SQL queries. It is widely used for data cleaning and transformation, log processing, real-time risk control, predictive analytics, and interactive queries. As an optimized alternative to Hadoop ecosystems, it is also one of the most popular computing frameworks in the current big data domain.

3.4 Data storage technology

HDFS (Hadoop Distributed File System), as the core storage system of the Hadoop ecosystem, is a distributed storage system specifically designed for large-scale data storage and distributed computing. It provides high

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

scalability through dynamically adding nodes and supports high-fault tolerance and high-throughput data access. HDFS adopts a Master-Slave master-slave architecture, where the master node NameNode manages file system metadata such as filenames, directory structures, and Broker locations, without storing actual data. Slave nodes DataNode store actual data blocks and periodically send heartbeats and block reports to NameNode, allowing the master node to dynamically monitor real-time data storage status(Wu,2021). Additionally, there are auxiliary nodes Secondary NameNode that regularly merge fsimage and edit files to reduce NameNode recovery time. The architecture diagram is shown in Figure 2.

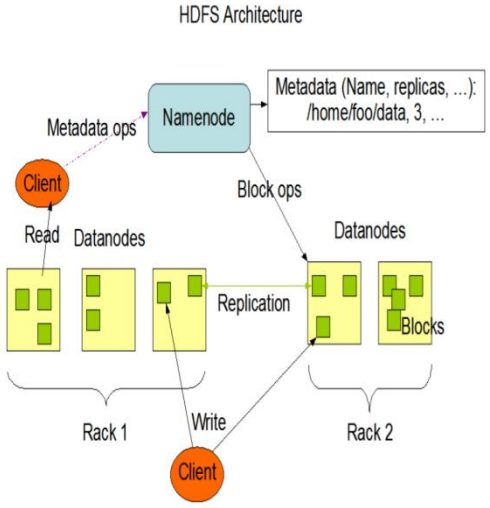

Figure 2. HDFS Architecture

When data is written, the client requests the NameNode to write the file, and the NameNode returns the available DataNode list. After the client divides the data into blocks and writes the data to the DataNode in Pipeline mode, the DataNode accepts the data and replicates it until the number of copies is satisfied. When reading data, the client requests the file block location from NameNode, and NameNode returns the nearest DataNode list. The client directly reads the data from the DataNode in parallel.

3.5 Data analysis model

This is a machine learning algorithm based on integrated learning (Ensemble Learning), composed of multiple decision trees (Decision Trees). It enhances the accuracy and stability of predictions through "voting (classification)" or "averaging (regression)." The key parameters for its analysis include n_estimators (the number of trees in the forest), max_depth (the maximum depth of a single tree), max_features (the maximum number of features considered during splits), and min_samples_split (the minimum number of samples required for a split). During model algorithm training, N data points are drawn back into the original dataset with replacement using bootstrap sampling (Bootstrap Sampling), generating multiple training subsets. Different decision trees are then trained on each subset. When splitting each tree, only a randomly selected portion of the features is considered (rather than all), to increase diversity and reduce overfitting. Finally, the prediction result is determined by the voting (classification) or averaging (regression) of all trees(Yu et al,2021).

In the mathematical derivation of the analysis model, if the problem to be processed is classification, the majority voting mechanism is used to aggregate the prediction results of all decision trees.

$$y = \text{mode}(\{h_1(x), h_2(x), ..., \}, h_T(x)\})$$

In the formula:
y : the final predicted category of sample X;
Hi (x): The prediction result of the i-th tree for sample x;
T: The total number of decision trees in the random forest;

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In addition, when performing classified voting, feature importance calculation is required to evaluate the importance of each feature by calculating the mean reduction in impurity caused by the split of each feature in the tree.

$$\text{Importance}(f) = \frac{1}{T}\sum_{t=1}^{T}(\text{ReductioninImpuritybyfintreet})$$

If the problem is regression (continuous predictions), the random forest will use the mean rather than the mode.

$$y = \frac{1}{T}\sum_{i=1}^{T} h_i(x)$$

Random forest realizes classification prediction by voting majority, because a single decision tree is easy to overfit, while multiple trees can vote to smooth noise. With the increase of the number of trees T, incorrect predictions will be covered by most correct predictions, reflecting the core idea of ensemble learning that collective decision is better than individual.

3.6 Visualization technology

The visualization module is a core component of data analysis, machine learning, and business intelligence systems. Its function is to transform complex data into intuitive graphics or interactive interfaces, helping users understand data patterns, discover trends, and support decision-making(Zhang,2021). Through systematic design of visualization modules, the efficiency of data-driven decision-making can be significantly improved. In practical construction, a balance must be struck between aesthetics, functionality, and performance.

ECharts is an open-source data visualization library based on JavaScript, enabling the rapid creation of interactive charts. It offers a rich variety of chart types, including line graphs, bar charts, pie charts, and scatter plots, with advanced customization for themes, interactions, and animations. The library supports multiple platforms, such as PCs and mobile devices, and can meet complex requirements through community plugins. Therefore, it is widely used in data analysis and large screen presentations.

4. SYSTEM DESIGN

Combining the characteristics of coal mine gas explosions with big data technology, a coal mine gas explosion early warning system is constructed through multi-source data fusion, real-time analysis, and intelligent alerts. This system achieves real-time collection of multidimensional data such as gas concentration, wind speed, and temperature. Historical data sets are used for model training, while real-time data is employed for explosion risk prediction. The system also features a visualization module for large screen display and finally triggers email alerts.

4.1 System architecture

Using Hadoop to build a data collection, mining, and analysis system specifically for predicting coal seam and gas outburst risks requires covering multiple steps from data collection to real-time analysis, historical data comparison, and risk prediction. According to actual application needs, the system architecture will include data collection layer, data storage layer, data procession layer， risk prediction layer, and visualization and alarm layers, as shown in Figure 3.

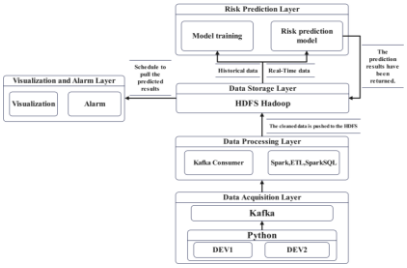

Figure 3. System Architecture

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The data collection layer collects real-time sensor data from coal mines (such as gas concentration, pressure, humidity, and temperature) through IoT devices, which is then sent to the data processing system via Kafka or other message middleware. The data storage layer uses a Hadoop HDFS distributed storage system to store both raw and cleaned data. In the data processing layer, after entering HDFS, the data undergoes cleaning, analysis, mining, and feature engineering using Spark, analyzing correlations between different data points, identifying key factors that may lead to coal and gas outbursts, aggregating sensor data, such as average gas concentration and temperature within time windows, and analyzing for any abnormal fluctuations. The risk prediction layer trains models using historical data, and analyzes real-time data with trained models (such as regression, classification, and time series) to predict coal seam and gas outburst risks. The visualization and alarm layer uses a JavaScript-based visualization database ECgarts to display data on gas concentration, temperature, humidity, and risk predictions in coal mines, importing the prediction results into these tools to generate visual dashboards. When indicators exceed thresholds, an alarm system (such as email or SMS notifications) is set up to issue warnings.

4.2 Data collection and processing

The system uses Apache Kafka as the data acquisition tool, connecting it to sensors deployed underground in coal mines to collect daily production data in real time. At the production system end, sensors act as data receivers, receiving and parsing the daily production figures from the mine, and creating SQL statements for adding, deleting, or modifying records in the main database, thereby achieving the purpose of restoring data records and status in the main database. At the real-time data analysis system end, sensors serve as data publishers, parsing the records received from the main database and publishing them in real time through the Kafka producer client.

In the coal mine safety production information collection system, there are abnormal values, missing values, disordered timestamps, and physical constraint conflicts in the collected information. In order to improve the quality and reliability of the original data to support accurate decision-making and analysis, it is necessary to preprocess the collected raw data, eliminate noise and anomalies, fill in missing information, and perform data fusion.

4.2.1 Data noise elimination

In the data processing of coal mine production, noise and outliers will seriously affect the accuracy of data analysis. SPARK (as a distributed computing framework) can efficiently process large-scale data and combine mathematical methods to eliminate noise and outliers.

Data noise is typically caused by random fluctuations or measurement errors. For coal mine production data noise, the exponential smoothing method is commonly used to eliminate noise. Let $\hat{x}_t$ be the smoothing value at the current moment, given a time series data $\{\hat{x}_t\}$, t=1,2,...,n, with a smoothing coefficient $\alpha \in (0,1)$, then its value is

$$\hat{x}_t = \alpha x_t + (1-\alpha)\hat{x}_{t-1}$$

In the formula:

$\hat{x}_t$: The smooth value at the current moment.

x_t: The original value at the current moment.

$\hat{x}_{t-1}$: The smooth value of the previous moment.

In the mathematical derivation process, it is known that the closer α is to 1, the more sensitive it is to recent data, and the smoother effect becomes weaker. The closer α is to 0, the higher the weight of historical data, and the smoother effect becomes stronger. Considering the characteristics of coal mine data, when eliminating noise, α is typically chosen within the range (0.1, 0.3) based on the degree of noise. In the system environment setup, Apache Spark is used to clean and process the data. Since EWMA (Exponential Weighted Moving Average) requires iterative calculations in chronological order, this needs to be implemented using the pandas_udf function. The specific code implementation is as follows:

```
from pyspark.sql.functions import pandas_udf
```

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

```
import pandas as pd
# Define the exponential smoothing function
@ pandas_udf("double")
def exponential_smoothing(series: pd.Series, alpha: float) -> pd.Series:
   return series.ewm(alpha=alpha, adjust=False).mean()
# Apply smoothing (assuming data is sorted by time)
df = df.withColumn("smoothed_value", exponential_smoothing(df["value"], alpha=0.2))
```

4.2.2 Data missing information filling

Coal mine production data often suffers from missing values due to sensor failures, communication interruptions, and recording omissions. To ensure the accuracy and reliability of subsequent data analysis and predictive results, it is necessary to fill in missing information to improve data quality. In Apache Spark, various methods for filling missing data are provided, such as mean and median imputation, linear regression imputation, KNN imputation, and matrix factorization imputation.

Since coal mine production data usually has strong temporal correlation, time series model can be added to fill in missing information by using matrix decomposition filling method. Let R^{m×n} be the original data matrix, then

$$\min_{U,V} ||W \odot (D - UV^T)||^2_F + \lambda(||U||^2_F + ||V||^2_F)$$

In the formula:

D ∈ R^{m×n} is a matrix of raw data containing missing values.

U ∈ R^{m×k} and V ∈ RΛ{n×k} are low rank decomposition factor matrices (k ≪ min (m, n)).

W ∈ {0,1}^{m×n} is the weight matrix (missing position is 0, observed position is 1).

⊙ Represents Hadamard product (element by element multiplication)

|| ||_F denotes the Frobenius norm λ> 0 is the regularization parameter

In the formula, in the data fitting term $||W \odot (D-UV^T)||^2_F$, only the difference between the observed data (the position with weight W=1) and the decomposition result UV^T is considered; missing positions (W=0) do not affect the objective function. The regularization term $\lambda(||U||^2_F + ||V||^2_F)$ primarily aims to prevent overfitting of the dataset, control the norms of matrices U and V, and ensure that the problem has a unique solution. During its solution process, the alternating least squares method is required. First, matrix V is fixed, making the optimization problem of the objective function with respect to matrix U.

$$\min_U ||W_U \odot (D - UV^T)||^2_F + \lambda||U||^2_F$$

W_U is the row weight of D. Then fix the matrix U so that the objective function for the optimization problem of the matrix V is

$$\min_V ||W_V \odot (D - UV^T)||^2_F + \lambda||V||^2_F$$

W_V is the column weight of D. Finally, it is decomposed into n independent regression problems for solving. After obtaining the optimal matrix U and V, the missing values can be filled in

$$\hat{D}_{ij} = (UV^T)_{ij}$$

In Apache Spark, the ALS algorithm can be used to implement the alignment, and the specific code is as follows:

```
from pyspark.ml.recommendation import ALS
als = ALS(rank=k, maxIter=10, regParam=λ,
```

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

```
    userCol="row_idx", itemCol="col_idx",
    ratingCol="value", implicitPrefs=False,
    coldStartStrategy="drop")  # Handling of missing values
model = als.fit(data)
complete_matrix = model.transform(full_index_pairs)  # Contains pairs of all locations
```

4.2.3 Data fusion

Based on the aforementioned data processing, big data technology is employed to integrate and process the data. The K-means clustering algorithm is primarily used for accurate information collection(Zhong et al,2021). Assuming that the given coal mine monitoring system has n sensors, the coal mine production dataset is X= {x1, x2,..., xn}, where xi ∈ Rd. Each data point xi represents d-dimensional production data at time t (such as gas concentration, temperature, vibration, etc.). The goal is to select the optimal sampling strategy, with the K-means clustering objective function being

$$\mathcal{J}_{\text{uncertainty}}(X_S) = \sum_{x \in X_S} \min_k || |x - \mu_k||^2$$

In the formula:

XS:Sample data subset.

μk：The centroid of the kth cluster

In order to obtain safety production information, a higher sampling rate of key data is required in the data collection of coal mine sensors. Therefore, dynamic weight is defined as

$$w(x_t) = \alpha \cdot \text{uncertainty}(x_t) + \beta \cdot \text{anomaly}(x_t)$$

In the formula:

$$\text{uncertainty}(x_t) = \min_k || |x_t - \mu_k||^2$$

$$\text{anomaly}(x_t) = ||x_t - \mu_{k(x_t)}||/\sigma_{k(x_t)}$$

After the above calculation, the objective function of information sampling is obtained

$$\mathcal{J}_{\text{dynamic}}(\text{X}_\text{S}) = \sum_{\text{x} \in \text{X}_\text{S}} \text{w (x)}$$

Based on the above process, noise processing, missing information filling and data fusion are carried out to ensure the quality and credibility of the data collected by the data collection layer, so as to ensure that the data can be used for subsequent data analysis and prediction.

4.3 Data storage

This system adopts the Hadoop HDFS distributed file storage system. Considering storage costs and preventing single points of failure in storage disks, the deployment does not involve setting up HDFS Hadoop on local clients but instead uses WebHdfsService on Alibaba Cloud servers to deploy HDFS in the cloud. During data storage, HDFS operations are executed via HTTP/REST API, which reduces resource consumption for data management and enhances system performance.

After injecting the URL of WebHdfsService into the configuration file, a PUT request needs to be sent to the op=MKDIRS interface, and the Boolean field in the returned JSON data should be parsed to create HDFS directories: tmp, ods, and ads. The tmp folder is used for storing temporary data collected by Kafka, the ods folder stores raw data after Spark cleaning, and the ads folder is used for storing data files with prediction labels and confidence levels.

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

After the directory is successfully created, you need to specify the method for file upload. In this system, the file upload process is divided into two stages. The first stage involves sending a request to NameNode in HDFS to obtain an DataNode address. Once the DataNode address is successfully obtained, the process moves on to the second stage. During this stage, a direct connection can be established with the DataNode to transmit data, allowing any Java object to be transmitted and automatically converted to JSON format.

Similar to the file upload method, the file download method also includes two stages: first, send a GET request to NameNode to obtain the redirect URL of DataNode, and then directly establish a link with DataNode to obtain data and save it locally. After defining the methods for file upload and download, it is necessary to define the management rules for files. In this module, the rules for file deletion are specified as follows: use a DELETE request to delete files or directories, with recursive=true indicating recursive deletion, and return whether the operation was successful at the end of the program. For moving and renaming files, use a PUT request to rename or move files, and specify the target path using the destination parameter.

The final step is to organize the list of stored files and read the file content. In the file listing function, it is necessary to obtain the list of files in the storage directory and return the name of the file list, then list the directory contents. In the file reading function, a GET request must be sent to obtain the redirect URL, and then read the file content from the redirect URL and return the file content string.

4.4 Data analysis

The system employs Apache Spark MLlib and Spring's random forest classification to achieve real-time analysis of daily production event stream data in coal mines, and uses pre-trained models to predict new data. During the model training process, the training code needs to use training data for feature engineering and logical training. After training is complete, the accuracy of the model must be evaluated, and the model along with the standardizer should be saved.

4.4.1 Model training

For the construction of training models, in combination with the unique environmental conditions of coal mines, data preparation and feature engineering must be conducted first. This involves integrating real-time sensor data from the mine (such as gas concentration, wind speed, vibration) and geological ledger data (coal seam thickness, gas content) to generate production log data. Time series data fusion techniques are employed, setting time windows for data correlation matching. For different sampling frequency data sources (such as 1Hz gas sensors and 0.1Hz temperature sensors), linear interpolation is used to align frequencies. Nonlinear composite features are constructed from multi-source data in coal mines, utilizing logarithmic multiplication for feature engineering processing.

In terms of model optimization, nearly one month's data from coal mines (a total of 19,991 data points, each including coal seam fracture density, gas concentration, gas pressure, rock stress, gas extraction volume, microseismic events, and temperature) was used for training. Before training, the warning samples were non-uniformly processed with normal samples, adjusting their ratio to 1:3 to optimize the data sample.

In order to make the training effect of the model reach the expected, according to the existing hardware conditions, the number of trees in the random forest model is set to 100, the maximum depth is 10, the minimum sample size for splitting is 5, and the feature sampling ratio is 0.5. In order to enhance the stability, bootstrap sampling is closed in high gas areas.

```
param_grid = {
  'n_estimators': [100],
  'max_depth': [10],
  'min_samples_split': [ 5],
  'max_features': ['sqrt', 0.5],
```

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

```
    'class_weight': [{0:1, 1:3}, {0:1, 1:5}],
    'bootstrap': [True, False]
}
```

The "early recall rate" index is customized. For the samples of actual accidents, the model is considered to make a correct prediction if it issues an early warning within 30 minutes before the accident. Compared with the traditional recall rate, this index is more in line with the safety requirements of coal mines.

```
from sklearn.metrics import make_scorer
def early_recall_score(y_true, y_pred):
    tp = np.sum((y_pred == 1) & (y_true == 1))
    fn = np.sum((y_pred == 0) & (y_true == 1))
    return tp / (tp + fn + 1e-5)
scorer = make_scorer(early_recall_score, greater_is_better=True)
```

The SHAP value analysis technology was used to interpret the characteristic contribution of high-risk samples (prediction probability> 70%), and to rank the key causes and analyze the influence of geological conditions related to coal mine gas explosion.

```
high_risk = X_test[(y_test == 1) & (model.predict_proba(X_test)[:,1] > 0.7)]
shap_values = explainer.shap_values(high_risk)
shap.summary_plot(shap_values[1], high_risk, plot_type='bar', max_display=10)
```

4.4.2 Risk profile

After completing the data analysis model training, it is necessary to create a random forest classification prediction service based on Apache Spark MLlib and Spring. This service will be used to load pre-trained models and analyze new data for predictions. When building the prediction model, attention should be paid to the data format imported, with column order consistent with the training phase. Define parameters for reading data types, adjust input parameters to List<List<Double>> data, indicating multiple sets of feature data. Multiple feature columns need to be merged into one vector and standardized using the pre-trained standardizer. The output parameter is Map<String, List<Double>>, containing the predicted label labels and confidence level confidences.

After data processing is complete, use the pre-trained random forest model for prediction and extract the prediction results. The predicted category is either 0 or 1, with probability distributions following vector types. The confidence value is determined according to the following rules: if the label ≤ 0.1, take the probability of category 0 (probability [0]); otherwise, take the probability of category 1 (probability [1]). After the program runs, it returns a list containing the predicted categories and their corresponding confidence levels.

4.5 Visualization and Alarm

This system's large screen visualization function is implemented using a JavaScript-based ECharts. According to actual needs, the data update mechanism is set as real-time data maintaining a long connection through WebSocket, ensuring that warning information is transmitted from edge devices to the screen within 800ms. Historical trend data uses a polling mechanism, with compressed data packets being retrieved from the cloud database every 3 minutes. Warning messages are pushed to achieve sensor marking and real-time status alerts.

The implementation of the alarm function is based on the Springboot-mail component, constructing a mail sending tool class to send both plain text and HTML formatted emails. It realizes the prediction and subsequent warning information. In the program, set the sender, recipient, subject, and body text, check for new high-risk logs

Research Article

predicted, and if they exist, send an alarm email. The email content includes device ID, submission time, predicted value, confidence level, and various sensor data.

5. RESULTS and discussion

This paper addresses the characteristics of diversified and heterogeneous data sources in coal mine safety systems. Combining big data technology, it studies and designs a coal mine gas explosion early warning system based on data mining techniques, mathematical derivation, coal mine safety management, and big data analysis methods. The system features data collection, data processing, data storage, risk prediction, visual presentation, and email alerts.

In terms of data collection and processing, Apache Kafka is used as the data collection tool to achieve the collection and publication of large amounts of daily production data from coal mines. Apache Spark is employed to eliminate noise from the raw data using exponential smoothing, and missing information is filled in by matrix factorization incorporating time series. Then, K-means clustering algorithm is used for data fusion to align the sampling rates of data at different frequencies. This enhances the authenticity and validity of the data, ensuring the scientific and accurate nature of data analysis.

In terms of data analysis, the random forest model is used to train and analyze the data for prediction. Its anti-overfitting and high accuracy make the predicted labels and confidence of coal mine gas explosion risk prediction credible and accurate.

In terms of data storage, the high fault tolerance and scalability of HDFS distributed file storage system ensure the security of data storage. Its copy mechanism avoids data loss caused by node failure, and can be infinitely expanded. The storage capacity and computing capacity increase linearly with the increase of nodes.

In terms of visualization and alarm, the rich interface provided by ECharts enables security decision makers to intuitively observe the dynamic changes of data and the location of risk points, and provides real-time alerts through email to ensure the real-time nature of security decisions.

To sum up, the coal mine gas early warning system based on big data can effectively improve the risk prevention and control ability and decision-making efficiency of coal mine gas explosion.

REFERENCES

[1] Li Guoqing, Li Xueyu, Hou Jie, Qiangxingbang, Wang Hao, Guozhen Xiang & Zhao Wei.(2022). Research and development of big data analysis system for mine safety hidden hazard identification and early warning. Metal Mine (06), 129-137.
[2] Li Xin.(2022). Design of Embedded Coal Mine Safety Production Information Acquisition System Based on Big Data. Coal Technology (06), 185-188.
[3] Dong yuan.(2019).Research on the susceptibility assessment of geological disasters in the big data environment (doctoral dissertation, China University of Geosciences).
[4] Gao Jing, Zhao Lijun & Lu Xuyang.(2022). Big data platform for coal mine safety management based on data mining. Coal Mine Safety (06), 121-125. doi: 10.
[5] Guo Wenhao.(2019). Establishment and applied research of geological disaster big data mining framework based on Hadoop (master's thesis, Kunming University of Science and Technology).
[6] He Wenna.(2013). Research on geological informatization based on the Internet of Things and cloud computing in the era of big data (doctoral dissertation, Jilin University).
[7] Jane Sharp.(2019). The Development and Key Technology of Geological Disaster Information Management System Based on Big Data. Jingwei World (02), 3-5 + 7.
[8] Liu Junqi, Liu Qiang, Liu Qianhui, Zhang Xialin, Lin Chen, Zhou Xin & Li Guoce.(2021). Discussion on Geological Disaster Data Management and Application Mode in the Big Data Era. Geological Science and Technology Bulletin (06), 276-282 + 292.
[9] Cheng Xiaolu & Yu Ningyu.(2021). Study on Critical Rainfall Conditions of Geological Disaster Classification Based on Meteorological Big Data.Shaanxi Geology (02), 71-80.

Journal of Information Systems Engineering and Management
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[10] Li Peng 16.(2022). Application of Information Technology Based on Internet of Things in Geological Disaster Monitoring and Early Warning. Value Engineering (16), 159-162.
[11] Pan Junfeng, Feng Meihua, Lu Zhenlong, Xia Yongxue, Xu Gang, Ma Hongyuan...& Zhang Jian.(2021). Research and application of comprehensive monitoring and early warning platform. Coal Science and Technology (06), 32-41.
[12] Pei Zhongmin, Li Bo, Xu Shuo & Zhu Hua.(2012). Integrated platform architecture of coal mine Internet of Things based on cloud computing. Coal Science and Technology (09), 90-94.
[13] Qiao Wei, Jin Dewu, Wang Hao, Zhao Chunonghu & Duan Jianhua.(2020). Construction of intelligent early warning platform for coal mine water damage monitoring based on cloud service. Coal Journal (07), 2619-2627.
[14] Wang Jie, Wang Chunhua, Li Xiaohua & Kelissa Yu.(2021). Design and research of big Data analysis cloud platform for coal Industry. Coal Engineering (09), 187-192.
[15] Wu Xiayun.(2021). Design and Realization of Guizhou Slope Geological Disaster Monitoring System based on Big Data (Master's thesis, Guizhou University).
[16] Yu Guofeng, Yuan Liang, Ren Bo, Li Lianchong, Cheng Guanwen, Han Yunchun...& Ma Jiguo.(2021). Big data prediction and early warning platform for water disaster on the bottom board. Coal Journal (11), 3502-3514.
[17] Zhang Wenhui.(2021). Research on geological disaster early warning model based on big data (Master's thesis, China University of Geosciences (Beijing)).
[18] Zhong xiaoxing, Wang Jiantao & Zhou Kun.(2021). Research Status and Intelligent Development Trend of Mine Coal Spontaneous combustion Monitoring and Early Warning Technology. Industrial and Mining Automation (09), 7-17.