Random Forest-Based Decision Tree Framework for Hazard Management in University Laboratories

Yunan Zhang*, Xiaoyu Wang
[1]National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, Beijing China
E-mail: zhangyn85@163.com
*Corresponding author

Keywords: decision tree (DT), laboratory hazardous chemicals, RF-based DT algorithm, university laboratory works

Received: October 24, 2024

This study explores the use of Decision Tree (DT) algorithms for detecting potential hazards in laboratory operations by analyzing a synthetic dataset modeled on historical accident reports. The dataset simulates mishaps and near-miss incidents in university laboratories, incorporating detailed descriptions of behaviors and risks. Feature extraction techniques like Principal Component Analysis (PCA) are used to train and test DT models. The Random Forest-based Decision Tree (RF-based DT) model demonstrates superior performance compared to traditional methods. Implemented in Python, the model predicts chemical hazards with 92.3% accuracy, 93.1% precision, 93.5% F1 score, and 91.4% recall. These results confirm the model's reliability for risk detection and management in laboratory settings.

Povzetek: Predlagan je model odločanja na osnovi naključnega gozda (RF-DT) za upravljanje nevarnosti v univerzitetnih laboratorijih, ki zanesljivo identificira kemijske nevarnosti in izboljšuje laboratorijsko varnost.

1 Introduction

University-conducted innovative experiments can involve the use of hazardous chemicals or laboratory procedures [1]. Additionally, they could be involved in risky tasks including handling pyrophoric materials, inactivating infectious pathogens, moving large gas cylinders, and completing metalwork with machine tools which are highly probable to end in accidents and near-miss occurrences [2].

The complexity of laboratory safety management grows when there is a chance of fire, explosion, and other issues. It is risky to undertake laboratory research given the rise in accident frequency [3]. Many different types of laboratories, including biological, chemical, electrical, mechanical, and environmental labs, are frequently found in one university. Every kind of laboratory is made up of several rooms with various purposes. Additionally, every area has a variety of tools and apparatus. This implies that running the university laboratory presents significant difficulties [4]. In actuality, issues with management work account for the majority of laboratory accidents. Daily management tasks depend on the equipment's routine inspection and maintenance to guarantee the safety of the laboratory. Manual statistics are typically used in the administration of the laboratory apparatus [5]. The task of equipment management is difficult, time-consuming, and costly since there are so many different kinds of equipment.

Faculty personnel, researchers, graduate students, or students could be killed in an incident that occurs in a university laboratory while conducting a chemical experiment [6]. The number of laboratory accidents at universities worldwide is unknown, but after they happened, similar incidents occurred again at other colleges [7]. In these instances, there wasn't a significant paradigm shift or change in the way laboratories are safe. Furthermore, the colleges' efforts to work together to stop similar incidents have not advanced [8].

The unusual and extensive attention from industry and mainstream media, as well as from academic institutions around the country, complicated the campus's response efforts even further [9]. The incident had a major impact on everyone in school, especially those who work in laboratory safety and research, the Office of Sustainability, Health & Safety (EH&S) is responsible for safeguarding the health and safety of all individuals on campus. [10]. We utilize an RF-based DT method in this work, which can handle a diverse set of features and provide robust predictions even when some features might be less relevant. Next, by recursively dividing the data based on these vectors, the DT method makes use to estimate hazard probability to anticipate the dangers related to university laboratory work.

The following is the order of this paper: Related works are included in section 2, and techniques are covered in section 3. Section 4 presents the experiment's results. Section 5

concludes with a summary of the study and suggestions for more research.

2 Related works

The inherent hazards assessment and categorization (IHAC) approach was developed in the study [11] for use in university chemistry labs as a result of this investigation. They conducted a quantitative assessment of the inherent risks in chemical laboratory materials, apparatus, and procedures. Concurrently, dividing labs into many levels can lead to more focused safety management and accident avoidance.

Seven substances' chemical concentrations were used in research [12], to examine the fluctuations in the potential of Hydrogen (pH) and Electrical conductivity (EC). The seven compounds were known to produce chemical spills regularly in South Korea and were classified as accident-preparedness substances. Furthermore, they compared the changes in pH, EC, and statistics during the dilution procedure to determine the probability of recognizing unknown chemicals.

The process of developing Standard Operating Procedures (SOPs) in research [13] afforded the chance to ascertain the necessary conditions for reaction setup, recognize possible dangers, define the appropriate handling of undesired materials, and conduct a comprehensive risk assessment. Here, they offered recommendations for SOPs that have to be created for university research facilities as well as an example of an SOP for the Grignard reaction.

A technique for the prospective evaluation of chemicals using sorting-based multi-parameter and multi-criteria decision-making (MCDM) hazards to worker safety in the university lab was established in research [14]. It was advised that certain control measures should be implemented to lower the laboratory's safety risk. The technique was meant to become a main source of information for university danger analysts and adjust to the risk assessment of university laboratories.

The fuzzy Bayesian network (BN) method combined with the human factors assessment and categorization system for university labs (HFACS-UL) was suggested in the study [15], to evaluate the risky conduct in university laboratories. The primary risk factors were determined by applying the model to an inference analysis. To identify further preventative and control methods, meta-networks and important agents were also investigated.

To discover the implementation of certain semi-quantitative methods was assessed in research [16], to determine potential bias or variances caused by utilizing different ways for the same tasks for chemical risk assessment. They could overcome the discrepancies observed in the risk assessment by using two or more distinct semi-quantitative instruments for every working activity they need to evaluate. The tactic could allow workers' contact with chemicals to be reduced.

After an analysis of statistical information to provide a broad explanation of the traits of greater and more frequent accidents, research [17] estimated and evaluated the total risk of the hazardous chemical sector using the entropy weight technique. It examined how safety laws have evolved in China's hazardous chemical business and how the sector was expected to grow moving forward.

The analytical tool based on the software, hardware, environment, and liveware (SHEL) paradigm was utilized in the study [18], to examine reports from accident investigations about explosions at two universities' chemical labs. Global university communities must collaborate to develop methods for research and analysis, instructions for writing accident reports, and an information-sharing platform that would enable them to take advantage of the knowledge gathered from a range of incidents.

Hazardous chemical control was a vital component of campus laboratory safety management, as demonstrated through study [19] and subsequent investigation of remedies and countermeasures. Realizing the intrinsic protection of the university laboratory, the safety management method was completed with the construction of the basis for safety management for its whole life of hazardous substances.

An explosive accident at a university laboratory was thoroughly investigated in the study [20], to determine the primary reasons and enhance safety management. The findings suggested that the experimenters' lack of caution and poor safety knowledge were the primary causes of the accident. To successfully prevent these kinds of tragedies, experimenters and related technical managers need to receive more safety training. It would help to foster a positive safety culture inside the institution.

It focused on methods for deciphering and discovering the possible reasons behind the actions of the people implicated in mishaps in chemical laboratories in the study [21]. Reflections could be beneficial for a variety of stakeholders, including administrators, graduates, suppliers and producers of chemicals and lab equipment, managers, universities and colleges making investments in new or renovated chemical laboratories, and environment, safety, and health (ESH) professionals.

There were many employment contexts, where people were exposed to chemicals in the study [22], but the investigation and healthcare facilities have not received enough attention. It examined how research laboratory staff were exposed to hazardous chemicals at work, evaluated their knowledge and attitudes about chemical hazards, examined whether they followed the rules for handling chemicals safely, and examined the impact of various factors on the important outcomes.

Table 1: Summary table

Method	Accuracy	Key Features	Limitations of SOTA
Proposed RF-based DT	High	Utilizes Random Forest and Decision Tree algorithms to recursively divide data for hazard forecast; and PCA for feature extraction.	Able to manage noisy datasets, scalable, and efficient for hazard identification in university labs.
IHAC Methodology [11]	Moderate	Quantitative assessment of intrinsic hazards in university chemistry labs.	Does not use machine learning for hazard prediction. Restricted scalability and adaptability.
Chemical Risk Analysis [12]	Moderate	Evaluate pH and EC variations to evaluate chemical spill likelihood.	Concentrates on chemical properties, lacking broader applicability to other laboratory situations.
SOPs Formulation [13]	Moderate	Creates Standard Operating processes to handle chemical hazards and reactions.	No predictive model; does not evaluate historical incident data for hazard prediction.
MCDM-based Risk Assessment [14]	Moderate	Multi-criteria decision-making for evaluating chemical dangers.	Does not utilize machine learning to forecast hazards. Risk evaluation is not automated or scalable.
Fuzzy Bayesian Network (BN) [15]	High	Fuzzy BN technique incorporated with human factors for hazard prediction.	Restricted to hazard behavior prediction and the absence of real-time scalability or practical incorporation.
Semi-quantitative Chemical Risk Assessment [16]	Moderate	Semi-quantitative techniques to detect biases in chemical risk evaluations.	May have inconsistent findings across differing situations and lacks an extensive predictive framework.
Entropy Weight Method for Chemical Risk [17]	Moderate	Utilizes entropy weight technique to quantify and assess risks in the chemical industry.	Concentrates on industry-level risk evaluation; not tailored to laboratory-particular incidents or hazards.
SHEL Framework for Explosion Analysis [18]	High	Utilizes SHEL framework for accident study and evaluation in chemical labs.	It absences predictive modeling capacities and does not tackle hazard probabilities in university labs.
Chemical Safety Regulation [19]	Moderate	Concentrates on regulating hazardous chemicals in university labs.	No predictive analytics or machine learning incorporation for hazard prediction.
Explosive Incident Analysis [20]	High	Examines causes of explosive incidents in university labs and suggests security enhancements.	Reactive method; concentrates on post-incident analysis rather than proactive hazard discovery.
Motivational Behavior Analysis [21]	Moderate	Examines motivations underlying individual activities in lab incidents.	Does not tackle the direct prediction of hazards or use automatic

			machine learning for risk evaluations.
Chemical Exposure Analysis [22]	Moderate	Evaluate research lab personnel's exposure to hazardous chemicals and assess safety protocol adherence.	Lack of predictive model for detecting hazard probabilities in real-time; no incorporation of historical data.

These previous studies in laboratory hazard prediction frequently absent predictive modeling, with most approaches being reactive or concentrating on particular risk factors like chemical properties or human behavior, rather than utilizing machine learning to predict hazards based on historical data. Additionally, numerous techniques lack scalability and may struggle to manage large datasets or real-time hazard detection, as evidenced by methods such as IHAC and the SHEL framework, which are restricted to particular situations and do not adapt well to the dynamic nature of laboratory work. Furthermore, these techniques struggle with noisy or incomplete datasets, which are common in practical uses, whereas the proposed RF-based DT algorithm can manage noisy data while still producing dependable predictions. Finally, while numerous SOTA techniques concentrate on isolated risk factors, the proposed approach incorporates multiple variables to create a more extensive and precise hazard prediction system, thus tackling the restrictions of previous techniques.

The suggested RF-based DT algorithm tackles the requirement for a more scalable, predictive, and comprehensive system for assessing and mitigating laboratory risks. It predicts risks by combining historical accident reports, feature extraction (PCA), and machine learning, providing a proactive solution for laboratory security management that is lacking in previous techniques. This gap in current research justifies the requirement for the RF-based DT algorithm, which can offer more precise hazard predictions while dealing with noisy, large datasets in real-time, providing an important improvement over the conventional approach.

3 Methodology

The collection includes recorded near-miss and accident events from a variety of academic laboratory operations. Benefits of Outliers, noisy data, and feature interactions can all be handled well by Random Forest. Several decision trees are constructed using Random Forest, each taking into account various combinations of these characteristics. Several decision trees are combined in Random Forest, an ensemble learning technique, to produce more accurate forecasts. We implement this approach in our suggested model. To anticipate the dangers connected with university laboratory work, the DT method uses these vectors to recursively segment the data based on their features, estimating hazard probabilities. The general flow is depicted in Figure 1.

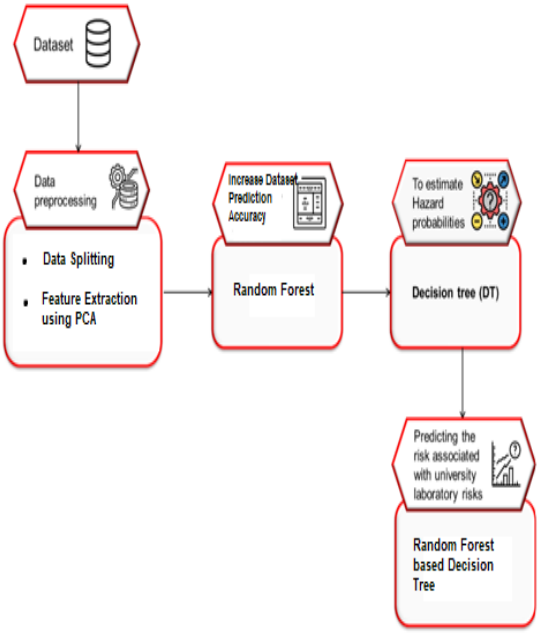

Figure 1: Overall flow

To enable replication, the methodology comprises data preprocessing, PCA for dimensionality reduction, and the creation of RF-based Decision Trees (RF-based DT). Preprocessing steps include dealing with missing values, encoding categorical variables, and normalizing numerical data. PCA is used to identify components that explain 95% of the variance. Random Forest is used to build decision trees, with cross-validation to improve hyperparameters such as tree count and depth. The trees are built by reducing impurities (such as the Gini index), and the RF-based DT combines them to improve model efficacy and generalization.

3.1 Dataset

A synthetic dataset consisting of event reports from university laboratory activities was created. Sample analysis, equipment maintenance, and chemical synthesis are a few of the laboratory tasks included in these reports. All of the reports consist of records on the dangers that would arise from those activities, inclusive of sample analysis, chemical synthesis, and equipment maintenance. The dataset accommodates 1133 reports which have been

randomly categorized as either close to pass-over occurrences or accidents to copy various levels of severity. This dataset is the foundation of research on DT algorithms to identify risks in lab operations. The dataset includes synthetic event reports that are intended to simulate laboratory behaviors and related hazards, such as sample analysis, chemical synthesis, and equipment maintenance. Each record indicates a hypothetical scenario based on practical laboratory functions, with detailed descriptions of behaviors, potential risks, and results. This synthetic method was selected due to the lack of extensive practical accident reports, while still retaining a realistic and controlled dataset for model assessment.

The synthetic dataset of 1133 reports were generated utilizing different laboratory operations, including sample analysis, equipment maintenance, and chemical synthesis, to represent common scenarios in academic laboratories. While the data is not based on actual accidents, it is designed to reflect plausible near-miss and accident scenarios that could happen in a university lab setting. The dataset was created by simulating various severity levels and includes features related to chemical procedures, safety protocols, and operational tasks. It has been pre-processed to guarantee consistency and usefulness for machine learning model training.

Data splitting: 20% percent of the input dataset is used for testing, while the remaining 80 % is used for training. A training dataset is a collection of data used to train a model. The testing dataset is also used to evaluate the performance of the trained model. The performance of each approach is created and evaluated using a variety of metrics, including accuracy, precision, recall, and F1 scores.

Feature extraction: Selecting the best features to extract is a crucial step because characteristics that aren't relevant could have a detrimental effect on the machine learning classifier's classification performance. In this step, principal component analysis, or PCA, is used to extract important features from the dataset.

3.2 PCA

Principal Component Analysis is a method for unlabelled feature extraction in data processing. Principal Component Analysis (PCA) was selected for feature extraction because it can decrease data dimensionality while retaining the most important variations in the dataset. PCA is especially helpful for dealing with large datasets because it assists remove irrelevant or redundant features, increasing model effectiveness and precision. Unlike t-SNE, which is mainly employed for visualization and may distort high-dimensional structures, and LDA, which is best suited for supervised learning tasks, PCA excels in unsupervised scenarios by capturing maximum variance without assuming class labels, rendering it ideal for the laboratory accident dataset.

A new, smaller feature space will be used to display features. The new features identified from the results of the PCA extraction are the features containing the most important data. The primary elements are obtained by optimizing the variance of the data. Since there are fewer additional dimensions (features) than starting characteristics, data visualization is feasible in a low-dimensional principal component space. Determine the dataset using each attribute in the manner shown below eq.1:

$$\bar{w}_i = \frac{1}{m}\sum_{j=1}^{m} \begin{matrix} w_{ji}, \ j = \\ 1,2,\ldots\ldots,m \end{matrix} \quad \begin{matrix} \\ i = 1,2,\ldots\ldots,m \end{matrix} \qquad (1)$$

Were,

n – No. of features,

w_{ji} - Data j-sample with i-feature,

m - Number data of sample, and

$\bar{w}_i$ - Data i-feature.

Calculate Φ with the following Equation 2,

$$\Phi = [\Phi_{ji}] = [w_{ji} - \bar{w}_i] \qquad (2)$$

Φ - Matrix of size $m \times n$.

Calculate the covariance matrix using the following Equation 3,

$$D = \frac{1}{m-1}\Phi^S\Phi \qquad (3)$$

D - Matrix of size $n \times n$.

Calculate the features of the D matrix by calculating the subsequent Equation 4.

$$Det\ (\lambda J - D) = 0 \qquad (4)$$

D - Covariance matrix, and

J - Identity matrix.

After that, calculate the subsequent equation 5 to determine the eigenvectors w that correspond to the features of the record λ,

$$(\lambda . J - D)w = 0 \qquad (5)$$

Form matrix w' Using the associated features after sorting the eigenvectors according to the eigenvalues, starting with the greatest. Calculate the principal components as follows in equation 6.

$$PC = \Phi w' \qquad (6)$$

PCA was used to decrease the dataset's dimensionality by converting original features into principal components that maintained 95% of their variance. Important features extracted using PCA comprise activity type, hazard severity, and equipment utilization. These elements were then utilized as inputs to decision tree algorithms, allowing for more precise forecasting of possible laboratory risks while reducing the effect of irrelevant or redundant data.

3.3 Random forest

A popular algorithm for machine learning one of the components of the supervised learning approach is the Random Forest Algorithm. It applies to machine learning problems that involve both regression and classification. The concept of supervised learning combines multiple classifiers to address a difficult problem and improve the model's performance, in its basis. Popular embedding learning is Random Forest could be used to manage hazardous chemicals in the lab by encoding chemical names and qualities into high-dimensional vectors. This makes it easier to do similarity and group chemicals according to their attributes. This model offers an effective way to compute vector representations using the Continuous Bag of Model (CBOM) and Skip-gram designs, which are both basic neural network models with a single hidden layer. Using backpropagation and stochastic descent of gradients, this approach first creates a PCA feature extract derived from the data entered. After that, the vectors are learned. The CBOM architecture predicts accuracy from future and previous contexts by utilizing a log-linear classifier learned through the negative collection and averaging contextual vectors. With a given word, the Skip-gram architecture predicts surrounding components. Training examples are created by eliminating a predetermined number of contextual phrases since the context is unbounded, such as $x_j - 3, x_j - 4, x_j + 3, x_j + 4$, thus the term "skip-gram." Figure 2 shows the CBOM and skip-gram structure.

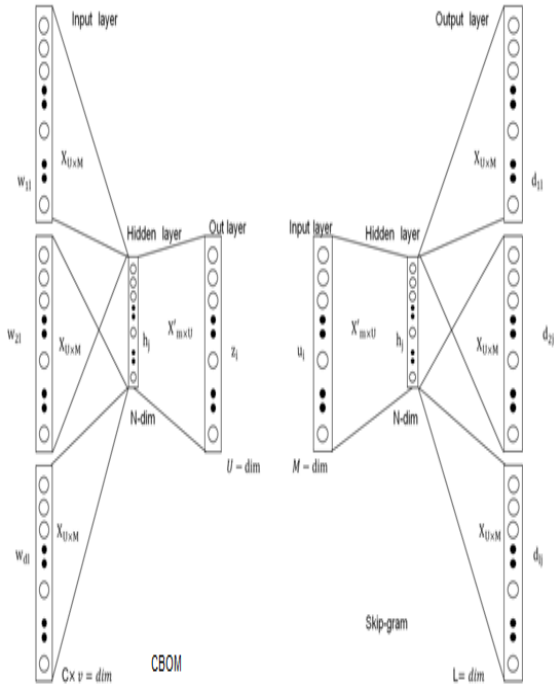

Figure 2: CBOM and skip-gram

Instead of creating a new model from scratch, the pre-trained model is utilized to address a comparable scenario or issue. Pre-trained models are developed and made available to the public for use in research. Pre-trained RF and Pre-trained GloVe are the two pre-trained models that are available for embedding models.

To improve the model's capacity to interpret various chemical and operational data, this paper employed embedding techniques like CBOW and Skip-gram, which are frequently employed for text analysis, to represent chemical names, properties, and lab-related descriptors as high-dimensional vectors. These embeddings allowed the model to compare and categorize features numerically, resulting in higher classification accuracy.

Before using Random Forest, Principal Component Analysis (PCA) was employed to decrease the dimensionality of the dataset by converting original features into principal components that maintained 95% of the variance. This decreased dataset was then embedded with CBOW and Skip-gram, allowing the Random Forest model to efficiently process the transformed data. The rationale for this method stems from embeddings' capacity to generate dense, numerical representations of categorical and text-like data, which improves model performance.

The Random Forest model improved its ability to classify hazards significantly by integrating PCA and embeddings, especially when evaluating reports of incidents involving chemical synthesis, sample evaluation, and equipment maintenance. The combination of these methods produced an effective and interpretable framework for hazard prediction in university laboratory settings.

3.4 Decision tree

The DT method uses these vectors to recursively split the data based on their features to determine hazard probability and anticipate the dangers related to the university chemical lab. The DT has branches containing qualities that determine the outcome, or the objective function and are organized in a sequential hierarchical structure. Nodes: arbitrary vertices where the potential course of events is ascertained, the outcome of Leaf (leaf) nodes with intends and values are used to depict the process of choosing a certain attribute value and merging several objects. Depending on the type of predicted indicator, decision trees could be divided into two categories: regression trees and classification trees. Trees of classification are useful for studying certain qualities, such as assigning items to a previously established class hence using them is advised when creating a prediction system. Data is categorized using decision trees, which split data into groups and provide a hierarchy of "if... then..." operators.
To separate the nodes into informative functions, create an objective function. Every division in which we optimize the increase is:

$$JH(C_o, e) = J(C_o) - \sum_{i=1}^{n} \frac{M_i}{M_o} J(C_i) \quad (7)$$

Where$e$ is the property that is used to conduct the splitting; Parents C_o and C_i is the $i-th$ child nodes, while J is a heterogeneity measure. M_i is the number of specimens contained in the $i-th$ child node, M_o is the overall number of values in the parent node.

We use binary decision trees for simplicity and to shrink the multidimensional search space. The child nodes $C_{left}$and C_{right} in our scenario are:

$$JH(C_o, e) = J(C_o) - \sum_{j=1}^{d} \frac{M_{left}}{M_o} J\left(C_{left}\right) - \frac{M_{right}}{M_o} J\left(C_{right}\right) \quad (8)$$

Where$e$ is the property that is used to conduct the splitting; the parent and $i-th$ child node databases are denoted by$J(C_o)$. J is a heterogeneity metric; The total number of samples in the parent node is represented by C_o, the number of samples in the child nodes that are in the $i-th$ child node is represented by M_o, the child nodes' numbers of patterns are represented by M_{left} and M_{right}.

Determination of entropy for all classes $o(j/s) \neq 0^2$that is not empty.

$$J_G(s) = -\sum_{j=1}^{d} o(j/s) log_2 o(j/s) \quad (9)$$

Where $o(j/s)$ is the percentage of samples including a single node s and the class.

Therefore, if every sample in a node is a member of the same class, the entropy is zero, and if the distribution of classes is uniform, the entropy is at its maximum.

A criterion that reduces the possibility of misdiagnosis is the Gini statistic for heterogeneity:

$$J_H(s) = -\sum_{j=1}^{d} o(j/s)(1 - o(j/s)) = 1 - \sum_{j=1}^{d} o(j/s)^2 \quad (10)$$

Where $K_H(s)$ is the Gini measure of heterogeneity and $o(j/s)$ is the proportion of samples that fall under a class and a single node.

Classification error is an additional metric for heterogeneity.

$$J_\varepsilon(s) = 1 - max\{o(j/s)\} \quad (11)$$

Where s is the single node and $o(j/s)$ is the proportion of samples that correspond to a class,$J_\varepsilon(s)$ is the classifier error.

Although this criterion is less susceptible to changes in the capacity of the classes at the nodes, it is appropriate for trimming trees but not for growing trees.
Decision Trees (DT) were used to divide laboratory hazard data, forecast risks, and evaluate hazard probabilities in university chemical laboratories. The objective function is critical in determining the best feature for dividing data at each node because it measures the efficiency of each split in terms of information gain or impurity reduction. The requirement for an objective function stem from the desire to determine the most pertinent features—such as chemical properties, laboratory activities, or prior incident history—that best differentiate hazardous and non-hazardous situations. By improving this function, the tree can recursively divide the data, guaranteeing that each child node indicates a more homogeneous subset of the data and thus enhancing the model's predictive accuracy. The goal is to reduce uncertainty regarding the risk of an incident at each split, resulting in better hazard forecasting. For instance, the objective function may prioritize divides that most efficiently distinguish incidents from non-incidents, allowing the decision tree to precisely forecast hazardous situations. Therefore, the objective function directly contributes to the primary goal of forecasting hazardous incidents by detecting important risk factors that influence decision-making at each node.

3.5 Random forest-based decision tree

An innovative technique for coping with hazardous chemicals in academic labs is the RF-based DT set of rules. This technique turns chemical descriptions and protection information into excessive-dimensional vectors that seize the complicated interactions between terms by using PCA and ML. The gadget can realize links among materials,

risks, and safety regulations for the reason that those vectors encode linguistic commonalities. The programs benefit in categorizing compounds in step with their traits and associated dangers using training a DT model on these vector representations. The hazardous chemicals machine gives more precise control and is streamlined by using this automatic class process, which reduces the need for human inspection and expertise. Furthermore, the machine gives extra particular hazard critiques and customized safety advice by using the semantic context under consideration. All things taken into consideration, there may be a great deal of capacity for elevating protection standards and decreasing dangers in university laboratories through the use of this novel technique. This suggests that Random Forest is a classifier that uses several decision trees on different dataset subsets and averages them to increase the dataset's predicted accuracy. The random forest predicts the result based on the majority of nodes of predictions from each decision tree rather than relying on just one. Accuracy is higher and overfitting is prevented because the forest has more trees.

The algorithm is known as Random Forest.
Step 1: Using Random Forest, n random lab records are chosen from the data set that contains k records from DT.
Step 2: A distinct decision tree is constructed for each sample.
Step 3: Every decision tree will generate an output of chemicals.
Step 4: Regression and classification averages are used, respectively, to assess the final product.

Pseudo code: RF-based DT Algorithm

Input:

Dataset: L = $\{(x_{1,}y_1), (x_{2,}y_2), (x_{3,}y_1) \dots \dots (x_{m,}y_m)\}$.

Laboratory set: A = $\{d_{1,}d_2 \dots \dots d_{a,}\}$

Method:

1. Build a node N.

2. If all samples in L belong to the same class or the features in L are uniform:

- Label N as a leaf node with the majority class in L.
- Return N.

3. Compute the best splitting node, d_n, from A utilizing the chemical selection Method.

4. Divide the dataset into subsets depending on the values of d_n.

5. For each subset corresponding to a value of d_n:

- If the subset is empty:
  - Append a leaf labeled with the majority class in L to node N.
- Otherwise:
  - Recursively call the TreeGenerate method with the subset and residual features to create child nodes.

6. Return node N.

Chemical Selection Method: Assess all features in A to determine the splitting node, d_n, that optimally partitions L into subsets utilizing a selection criterion (for example Gini index, information gain).

The RF-based DT algorithm utilizes a dataset with features representing hazardous chemical activities, and the chemical selection process concentrates on detecting the most important features for splitting data at each node. The splitting criteria are determined utilizing techniques that measure the node's heterogeneity, such as Gini impurity or entropy. The algorithm starts by choosing a random subset of the dataset and then creating individual decision trees with different subsets of attributes. The node-splitting procedure selects the most informative features for partitioning the data, to increase prediction accuracy and decrease overfitting.

The RF-based DT algorithm builds on the Random Forest methodology to handle hazardous materials in academic laboratories by utilizing laboratory incident reports that contain information on chemical properties, risk factors, and security guidelines. These reports are converted into high-dimensional vectors via PCA and machine learning, which encode the relationships between incidents, hazards, and security needs. The Random Forest framework generates numerous decision trees from random subsets of the incident data and then aggregates their forecasts to enhance accuracy and decrease overfitting. Unlike conventional Random Forests, this method includes semantic and contextual data from the vectors, allowing for more accurate chemical categorization and personalized security suggestions. By integrating the advantages of Random Forests and decision trees, the RF-based DT methodology improves hazard prediction and security standards, decreasing dependence on human oversight and reducing hazards in university laboratories.

4 Experimental results

The recommended method is implemented on a Windows 10 laptop with an Intel i7 core CPU and 8GB of RAM. To ensure the study's reproducibility, all software and libraries used are explicitly listed, along with their version numbers. Python 3.10.1 was used for analysis and model development, with important libraries including Scikit-Learn 1.2.2 for machine learning and preprocessing, TensorFlow 2.13.0 for sophisticated computations, and NumPy 1.24.3 for numerical operations. Data visualization was performed with Matplotlib 3.7.1 and Seaborn 0.12.2. The project was managed and carried out in a Jupyter Notebook environment, version 6.5.4. By documenting these particular versions, this study hopes to ease replication of the findings and guarantee consistency across various computational setups. The methods are Random Forest (RF), Decision Tree (DT), and RF-based DT (Proposed). The implementation used Python libraries like Scikit-Learn to create, train, and assess the models on the training and testing datasets. Cross-validation is used to optimize important hyperparameters like the number of trees (for Random Forest, 100 trees) and tree depth (for Decision Trees, the default depth is 5 trees).

The main goals of the research were to improve the accuracy of identifying hazardous chemical practices and conditions in university laboratories while guaranteeing accurate, dependable, and balanced detection using evaluation metrics like accuracy, precision, recall, and F1-score. Accuracy assesses the model's overall capacity to accurately classify hazardous and non-hazardous situations, whereas precision assesses the system's reliability in reducing false positives, or the incorrect detection of non-hazardous conditions as hazardous. Recall evaluates the model's capacity to identify true hazardous events while avoiding false negatives in which actual hazardous incidents are missed. Finally, the F1-score, calculated as the harmonic mean of precision and recall, offers a thorough assessment of the model's efficiency by balancing false positives and false negatives. Accuracy in the context of managing hazardous chemicals in university laboratories refers to a detection model's capacity to identify both safe operations and hazardous situations. It is a metric of the model performance overall in terms of accurately recognizing non-hazardous circumstances and forecasting real hazardous scenarios. Figure 3 and Table 1 display the accuracy performance. The methods are Random Forest (RF), Decision Tree (DT), and RF-based DT (Proposed), which achieve an accuracy of 85%, 88%, and 92.3%. Consequently, the RF-based DT is more accurate than the other methods used for Hazardous chemical handling at university laboratories. These metrics were selected to strike a balance between detection accuracy (precision) and hazard detection (recall). In lab security, the trade-off between precision and recall is crucial: high precision decreases false alarms, while high recall results in fewer missed hazards. The F1 score offers a balanced evaluation, rendering the model suitable for practical uses in which false positives and missed risks should be reduced.

Table 2: Values for precision, accuracy, recall, and F1 score

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Random Forest (RF)	85	88	84.2	82
Decision Tree (DT)	88	83.4	86.7	86
RF based DT (Proposed)	92.3	93.1	91.4	93.5

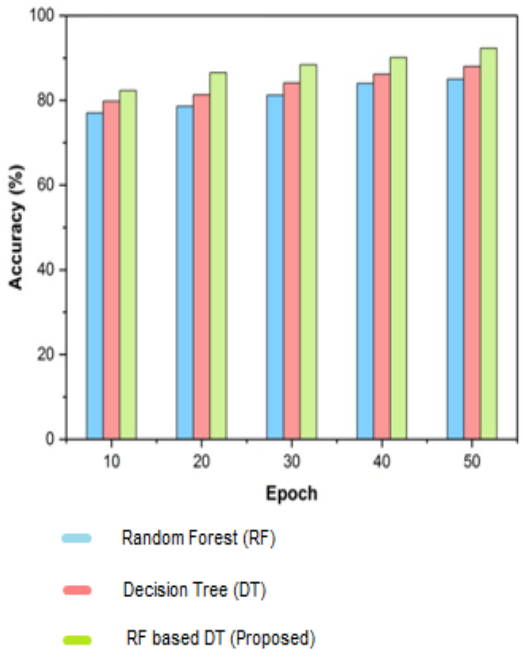

Figure 3: Accuracy Comparison of Random Forest, Decision Tree, and Proposed RF-based DT Models Across Epochs

Precision in university hazardous chemical management refers to the system's reliability in accurately recognizing hazardous chemical-related situations. In particular, it calculates the percentage of actual positive danger detections among all the systems detections.

The precision performance is shown in Figure 4 and Table 1. The approaches, which provide values of 93.1%, 88%, and 83.4%, are RF-based DT, Random Forest, and DT. As a result, while handling hazardous chemicals in university

laboratories, the RF-based DT is more precise than the other techniques.

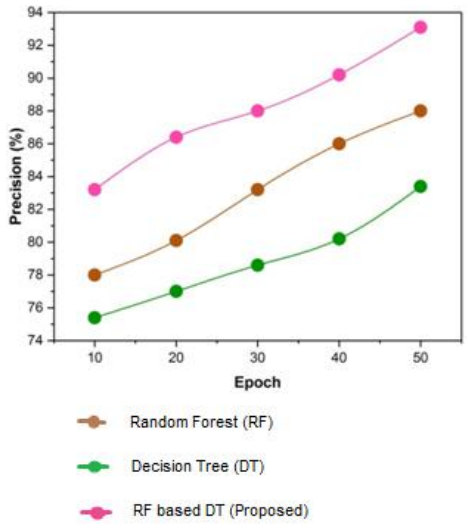

Figure 4: Precision comparison of random forest, decision tree, and proposed RF-based DT models across epochs

When discussing university laboratory hazardous chemicals management, recall refers to the statistic used to assess a detection model that identifies real hazardous situations. It can be defined as the proportion of properly classified hazardous occurrences on all actual hazardous incidents.

Recall performance is displayed in Table 1 and Figure 5. The methods RF-based DT, Random Forest, and DT offer values of 91.4%, 84.2%, and 86.7%. Therefore, the RF-based DT is more reliable than the other methods for handling hazardous substances in university laboratories.

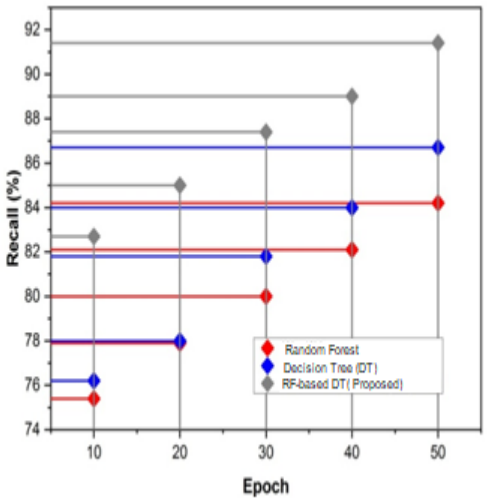

Figure 5: Recall Comparison of Random Forest, Decision Tree, and Proposed RF-based DT Models Across Epochs

An important indicator for assessing hazard detection models' performance in university laboratory chemical control is the F1 score. It is the precision and recall harmonic mean that balances false positives and false negatives.

The F1 score results are shown in Table 1 and Figure 6. Values of 93.5 %, 82%, and 86% are attained using RF-based DT, Random Forest, and DT techniques. Therefore, compared to other techniques used for managing hazardous chemicals in university laboratories, the RF-based DT is better.

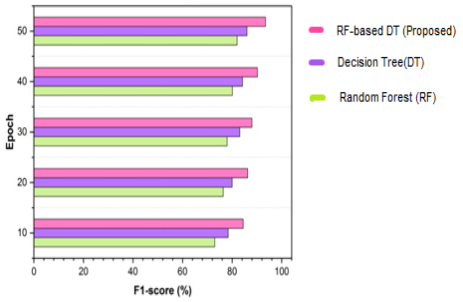

Figure 6: F1-score comparison of random forest, decision tree, and proposed RF-based DT models across epochs

The findings show that the proposed RF-based Decision Tree algorithm outperforms standalone Decision Trees and Random Forests for handling hazardous chemicals in university labs. Its excellent efficiency across all evaluation metrics demonstrates its resilience and practicality for real-world applications in hazardous chemical security. The RF-based DT achieves high values for accuracy, precision, recall, and F1-score, ensuring dependable detection of potential risks while decreasing the possibility of misclassification.

4.1 Discussion

The RF-based DT algorithm outperforms Random Forest (85%) and Decision Trees (88%) in terms of accuracy (92.3%) because it combines the advantages of ensemble learning and decision tree logic to capture intricate trends in hazardous chemical data, improving prediction precision. The utilization of Principal Component Analysis (PCA) for feature extraction substantially improves model resilience by decreasing dimensionality, reducing noise, and maintaining essential features, resulting in increased model stability and recall. However, this method has trade-offs, especially in terms of computational cost and scalability, as the combination of Random Forest and Decision Trees needs a significant amount of processing power, which may limit its application to larger datasets or real-time hazard detection systems. Despite these difficulties, the RF-based DT's better efficiency in accuracy, recall, and F1 score, combined with PCA's capacity to filter out irrelevant data,

make it a strong candidate for hazardous chemical handling, with optimization methods potentially tackling scalability issues.

This work expands on incremental efficiency gains by proposing an extensive framework for hazardous chemical management that integrates Random Forest (RF) and Decision Tree (DT) models with novel optimization methods and domain-specific tailoring. Unlike previous RF and DT executions, this method incorporates sensitivity analysis, hyperparameter tuning, and real-world data preprocessing tactics to tackle practical issues like imbalanced datasets, missing data, and domain-specific hazards. Additionally, the model's application in laboratory security provides a novel contextualization, with 92.3% accuracy resulting in actionable results like decreasing potential accidents and enhancing risk mitigation protocols. By aligning the methodology with real-world security enhancements, this study represents a significant advance in the area, bridging the gap between theoretical enhancements and practical executions.

Attaining 92.3% accuracy in hazardous chemical management is important because it shows that the model can forecast and classify potential risks related to laboratory operations. This level of efficiency leads to tangible enhancements in real-world laboratory safety by allowing for proactive measures like detecting high-risk chemicals, improving storage protocols, and enhancing staff training to prevent accidents. These improvements not only decrease the likelihood of incidents but also improve compliance with security requirements, resulting in a safer lab setting. However, the experimental setup has some constraints, especially the dataset size of 1133 reports, which, while adequate for initial validation, may limit generalizability to a variety of laboratory environments. Future research with larger, more diverse datasets is required to confirm the model's resilience and usefulness to various chemical processes and institutional settings.

The RF-based DT algorithm’s efficiency could be improved further by performing a sensitivity analysis to assess the influence of key hyperparameters like the number of trees in the forest, maximum tree depth, and minimal samples per split. This analysis would aid in determining optimal settings for increasing model precision and effectiveness in hazardous chemical management. Furthermore, the approach's resilience under adversarial conditions, like missing or mislabeled data, could be evaluated by performing controlled perturbations on the dataset and observing the model's effectiveness. Working with sensitive data, like laboratory accident reports, necessitates the highest ethical standards. This study protects data privacy by anonymizing datasets and adhering to pertinent data protection standards, encouraging the ethical and responsible utilization of machine learning in laboratory security management. These additional steps would offer greater insight into the model's dependability and social effect.

5 Conclusion

In this work, we used accident reports from before to examine DT algorithms to discover potential dangers in laboratory operations. The collection is made from near misses and accident reports from several instructional laboratory operations. In this study, we used accident reports from earlier to look at how nicely DT algorithms identify capability hazards in lab work. The collection was made up of near-miss and accident reports from a variety of academic laboratory operations. The nature of laboratory activities and associated hazards are included in each report. We used a RF-based DT technique in our suggested model, DT method uses these vectors to recursively split the data based on their features to estimate hazard probabilities and anticipate the dangers related to university laboratory work. When compared to the existing method, the proposed method achieves accuracy (92.3%), precision (93.1%), F1 score (93.5%), and recall (91.4%), respectively. Although the Decision Tree Algorithm simplifies the handling of hazardous chemical compounds in college laboratories, other developments in the vicinity want to consist of predictive analytics and non-stop surveillance for proactive risk discount and efficient use of sources.

References

[1] Ostad-Ali-Askari, K., 2022. Management of risks substances and sustainable development. *Applied Water Science*, 12(4), p.65. https://doi.org/10.1007/s13201-021-01562-7

[2] Gopalaswami, N. and Han, Z., 2020. Analysis of laboratory incident database. *Journal of Loss Prevention in the Process Industries*, 64, p.104027. https://doi.org/10.1016/j.jlp.2019.104027

[3] Wahab, N.A.A., Aqila, N.A., Isa, N., Husin, N.I., Zin, A.M., Mokhtar, M. and Mukhtar, N.M.A., 2021. A systematic review on hazard identification, risk assessment, and risk control in an academic laboratory. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 24(1), pp.47-62. 10.37934/araset.24.1.4762

[4] Nasrallah, I.M., El Kak, A.K., Ismaiil, L.A., Nasr, R.R. and Bawab, W.T., 2022. Prevalence of accident occurrence among scientific laboratory workers of the public university in Lebanon and the impact of safety measures. *Safety and health at work*, 13(2), pp.155-162. https://doi.org/10.1016/j.shaw.2022.02.001

[5] Harada, T., Hayashi, R. and Tomita, K., 2023. Prediction of hazards in laboratory work using deep learning models learned from past laboratory accidents. *Journal of Environment and Safety*, p.E23RP0601. https://doi.org/10.11162/daikankyo.E23RP0601

[6] Bai, M., Liu, Y., Qi, M., Roy, N., Shu, C.M., Khan, F. and Zhao, D., 2022. Current status, challenges, and future directions of university laboratory safety in China. *Journal of Loss Prevention in the Process Industries*, 74, p.104671. https://doi.org/10.1016/j.jlp.2021.104671

[7] Li, X., Zhang, L., Zhang, R., Yang, M. and Li, H., 2021. A semi-quantitative methodology for risk assessment of university chemical laboratory. *Journal of Loss Prevention in the Process Industries*, 72, p.104553. https://doi.org/10.1016/j.jlp.2021.104553

[8] Galasso, A., Luo, H. and Zhu, B., 2023. Laboratory safety and research productivity. *Research Policy*, 52(8), p.104827. https://doi.org/10.3386/w31313

[9] Ezenwa, S., Talpade, A.D., Ghanekar, P., Joshi, R., Devaraj, J., Ribeiro, F.H. and Mentzer, R., 2022. Toward improved safety culture in academic and industrial chemical laboratories: an assessment and recommendation of best practices. *ACS Chemical Health & Safety*, 29(2), pp.202-213. https://doi.org/10.1021/acs.chas.1c00064

[10] Schröder, I., Czornyj, E., Blayney, M.B., Wayne, N.L. and Merlic, C.A., 2020. Proceedings of the 2018 laboratory safety workshop: hazard and risk management in the laboratory. *ACS Chemical Health & Safety*, 27(2), pp.96-104. https://doi.org/10.1021/acs.chas.0c00012

[11] Liu, S., Ju, S., Meng, Y., Liu, Q. and Zhao, D., 2023. Inherent Hazards Assessment and Classification Method for University Chemical Laboratories in China. *ACS Chemical Health & Safety*, 30(4), pp.156-164. https://doi.org/10.1021/acs.chas.3c00022

[12] Nam, S.H., Ku, T.G., Park, Y.L., Kwon, J.H., Huh, D.S. and Kim, Y.D., 2022. Experimental study on the detection of hazardous chemicals using alternative sensors in the water environment. *Toxics*, 10(5), p.200. https://doi.org/10.3390/toxics10050200

[13] Chandra, T., Zebrowski, J.P., McClain, R. and Lenertz, L.Y., 2020. Generating standard operating procedures for the manipulation of hazardous chemicals in academic laboratories. *ACS Chemical Health & Safety*, 28(1), pp.19-24. https://doi.org/10.1021/acs.chas.0c00092

[14] Gul, M., Yucesan, M. and Karacahan, M.K., 2023. A Multi-parameter Occupational Safety Risk Assessment Model for Chemicals in the University Laboratories by an MCDM Sorting Method. In *Advances in Reliability, Failure and Risk Analysis* (pp. 131-149). *Singapore: Springer Nature Singapore*. https://doi.org/10.1007/978-981-19-9909-3_6

[15] Li, Z., Wang, X., Gong, S., Sun, N. and Tong, R., 2022. Risk assessment of unsafe behavior in university laboratories using the HFACS-UL and a fuzzy Bayesian network. *Journal of safety research*, 82, pp.13-27. https://doi.org/10.1016/j.jsr.2022.04.002

[16] Mastrantonio, R., Scatigna, M., D'Abramo, M., Martinez, V., Paoletti, A. and Fabiani, L., 2020. Experimental application of semi-quantitative methods for the assessment of occupational exposure to hazardous chemicals in research laboratories. *Risk Management and Healthcare Policy*, pp.1929-1937. https://doi.org/10.2147/rmhp.s248469

[17] Yang, D., Zheng, Y., Peng, K., Pan, L., Zheng, J., Xie, B. and Wang, B., 2022. Characteristics and statistical analysis of large and above hazardous chemical accidents in China from 2000 to 2020. *International journal of environmental research and public health*, 19(23), p.15603. https://doi.org/10.3390/ijerph192315603

[18] Fukuoka, K. and Furusho, M., 2022. A new approach for explosion accident prevention in chemical research laboratories at universities. *Scientific reports*, 12(1), p.3185. https://doi.org/10.1038/s41598-022-07099-2

[19] Yang, J., Xuan, S., Hu, Y., Liu, X., Bian, M., Chen, L., Lv, S., Wang, P., Li, R., Zhang, J. and Shu, C.M., 2022. The framework of safety management in a university laboratory. *Journal of Loss Prevention in the Process Industries*, 80, p.104871. https://doi.org/10.1016/j.jlp.2022.104871

[20] Chen, M., Wu, Y., Wang, K., Guo, H. and Ke, W., 2020. An explosion accident analysis of the laboratory in the university. *Process Safety Progress*, 39(4), p.e12150. https://doi.org/10.1002/prs.12150

[21] McLeod, R.W., 2022. Approaches to Understanding Human Behavior When Investigating Incidents in Academic Chemical Laboratories. *ACS Chemical Health & Safety*, 29(3), pp.263-279. https://doi.org/10.1021/acs.chas.2c00020

[22] Nasrallah, I.M., El Kak, A.K., Ismaiil, L.A., Nasr, R.R. and Bawab, W.T., 2022. Prevalence of accident occurrence among scientific laboratory workers of the

public university in Lebanon and the impact of safety measures. *Safety and health at work*, 13(2),pp.155-62. https://doi.org/10.1016/j.shaw.2022.02.001