APPENDICES

A RARE EVENT DATA

A.1 Naturally rare datasets: Table 3

Table 3. Naturally rare datasets.*

Sector	Event % & rarity group	Datasets with modality	Papers
Earth Sciences	0-1(R1)	meteorological (heatwaves) dataset(N,T), 3W dataset(N,T) [29]	[49, 187]
	1-5(R2)	Oil dataset(I, T) [113], 3W dataset(N,T) [29], Tornado dataset(N,T) [116]	[59, 113, 114], [129, 181, 187]
	5-10(R3)	seismic-bumps Data Set (N, T) [13], 3W dataset(N, T) [29]	[120, 187]
	10+(R4)	LIDAR data(I, T)[185], 3W dataset(N,T) [29]	[186, 187]
Manufacturing	0-1(R1)	pulp -and-paper dataset(N,T) [160], Bosch Production Line Performance dataset(N,T) [53]	[96, 160–162], [119, 133, 144, 202]
	10+(R4)	Case Western Reserve University Rolling Bearing dataset(N,T) [23], IMS bearing dataset)(N,T) [155], XJTU-SY datasets(N,T) [190], PRONOSTIA bearing dataset(N,T) [143]	[123, 165, 190, 203]
Telecommunication	0-1(R1)	probe binary-UCI(N, T), r2l binary-uci(N,T), KDD Cup 99 Data (N,T) [4, 5, 19, 164] Alarm data(n,T), VoIP traffic data(N, T) [33]	[127, 134, 195, 198]
Transportation	1-5(R2)	Air Pressure System(APS) Failure at Scania Trucks Data Set(N, T) [17, 20]	[73, 82, 157]
	5-10(R3)	MnDot traffic data(N, T) [33] Traffic Prediction Dataset(N, T) WWD Data (N, T) [26, 27]	[45, 134, 162]
Economy	0-1(R1)	S and P BSE SENSEX(N, T)[21, 99] Nifty 50(N, T) [24, 34] Kaggle Credit Card Fraud Detection(N, T)	[51, 145]
Healthcare	0-1(R1)	Thoracic surgery dataset(N, T) [121], Bioassay AID 746(N, T),687(N, T),456(N, T),373(N, T), Suicide data(N,T) [18]	[63, 121]
	1-5(R2)	stroke dataset(N, T) [22], Bioassay AID 362(N, T) [121]	[121, 147]
	5-10(R3)	Bioassay AID 1608(N, T) [121]	[121]
	10+(R4)	Wong's dataset from Canadian(TX ,T) [197]	[197, 211]
Energy	5-10(R3)	Daily electric energy production measurements dataset(Spain, 2003)(N,T)	[49]
Others	0-1(R1)	PCD dataset(I, T) [106],	[89]
	1-5(R2)	K1b-WebACE(N,T) [91]	[198]
	5-10(R3)	La12-TREC(N,T)	[198]
	10+(R4)	Recidivism dataset (N, T) , Audio-Anomaly-Dataset(A, T) [25]	[36, 98]

*N-Numeric, TX-Textual, I-Image, A-Audio, T-Time series

A.2 Derived datasets: Table 4

Table 4. Derived datasets.*

Sector	Event % & rarity group	Papers	Source Datasets with modality
Earth Sciences	0-1(R1)	[40, 60, 89]	ABCD (AIST Building Change Detection) dataset(I, T) [15, 16, 77], Weather data(N,T),Air pollutant data(N,T), Space Weather ANalytics for Solar Flares (SWAN-SF) benchmark dataset(N, T) [80]
	1-5(R2)	[40]	ABCD (AIST Building Change Detection) dataset(I, T) [15, 16, 77], Space Weather ANalytics for Solar Flares (SWAN-SF) benchmark dataset(N,T) [80]
	5-10(R3)	[60, 114]	Oil dataset(I,T) [113], Weather data(N, T), Air pollutant data(N,T)
	10+(R4)	[60, 139]	Weather data(N, T),Air pollutant data(N,T), ABCD (AIST Building Change Detection) dataset(I,T) [15, 16, 77],
Telecommunication	0-1(R1)	[95, 105, 198]	IEEE 39-bus power system data(N,T) [14], KDD Cup-99 (N,T) [4, 5, 19]
	1-5(R2)	[95]	IEEE 39-bus power system data(N, T) [14]
	5-10(R3)	[95]	IEEE 39-bus power system data(N, T) [14], Spam data(N, T) [6]
	10+(R4)	[95]	IEEE 39-bus power system data(N, T) [14], Spam data(N, T) [6]
Transportation	0-1(R1)	[41, 67]	AIRBUS data(N, T), ACMS dataset(N, T) [10]
Healthcare	0-1(R1)	[162]	EEG Seizure Dataset(N, T)
	1-5(R2)	[162, 200]	COVID-19(N, T), InP -Duke University Health System (DUHS)(N,T) [149], SEER(N, T) [167], EEG Seizure Dataset (N, T)
	5-10(R3)	[200]	COVID-19(N, T)
Energy	0-1(R1), 1-5(R2), 5-10(R3), 10+(R4)	[46]	MAGIC Gamma Telescope(N,T) [9]
Others	0-1(R1)	[89]	Augmented MNIST(I,T) [85], WDC dataset(I,T) [90]
	10+(R4)	[89, 120, 170]	Augmented MNIST(I,T) [85], Adult dataset(N,T) [3], AudioSet dataset(A,T) [84]

*N-Numeric, TX-Textual, I-Image, A-Audio, T-Time series

A.3 Simulated & Synthetic datasets: Table 5

Table 5. Simulated & Synthetic datasets.*

Sector	Event % & rarity group	Papers	Data type	Technique
Earth Sciences	0-1 (R1)	[132, 152, 187]	N, T	OLGA Dynamic Multiphase Flow Simulator [32], MATLAB
	1-5 (R2)	[132, 187]	N, T	OLGA Dynamic Multiphase Flow Simulator [32], MATLAB
	5-10 (R3)	[60, 114] [132, 187]	N, T	MATLAB, OLGA Dynamic Multiphase Flow Simulator [32],
	10+ (R4)	[132, 187]	N, T	OLGA Dynamic Multiphase Flow Simulator [32]
	Rarity not reported	[57, 58, 107]	N, T	Signal Fragment Assembler (SFA), Variational Autoencoder (VAE), Data Picker (DP), Quality Classifier (QC)
Others	0-1 (R1)	[46, 70, 105, 122, 146]	N, T	Monte Carlo, MATLAB
			N, T	Importance sampling
	1-5 (R1), 5-10 (R3)	[46, 70]	N, T	Monte Carlo, MATLAB
			N, T	Importance sampling
	10 + (R4)	[120]	N, T	MOA [52]

*N-Numeric, T-Time series

B DATA PROCESSING APPROACHES

B.1 Analyzing data cleaning approaches with data modalities, rarity groups, and downstream tasks

To explore the data cleaning approaches, we referred to 116 rare event prediction-related papers. Then, we analyzed these by data cleaning approach, modality, and rarity group. We observed the interplay between these as shown in Figure 7. In terms of numerical data cleaning, techniques such as data sifting, data filtering, imputation, and noise removal have been commonly employed, with the rarity level being independent of these techniques. Notably, noise removal has not been applied to the extremely-rare group. Moreover, specific modalities are addressed, such as image processing techniques for image-based rare event prediction and text-related techniques encompassing textual summary generation, text conversions, stemming, and lemmatization for text-based predictions. These audio-based and text-based approaches were used for frequently-rare datasets. Furthermore, image processing techniques are employed to predict images within the very-rare group. While many data processing methods have supported classification tasks, some of these methods have been used in clustering and forecasting research.

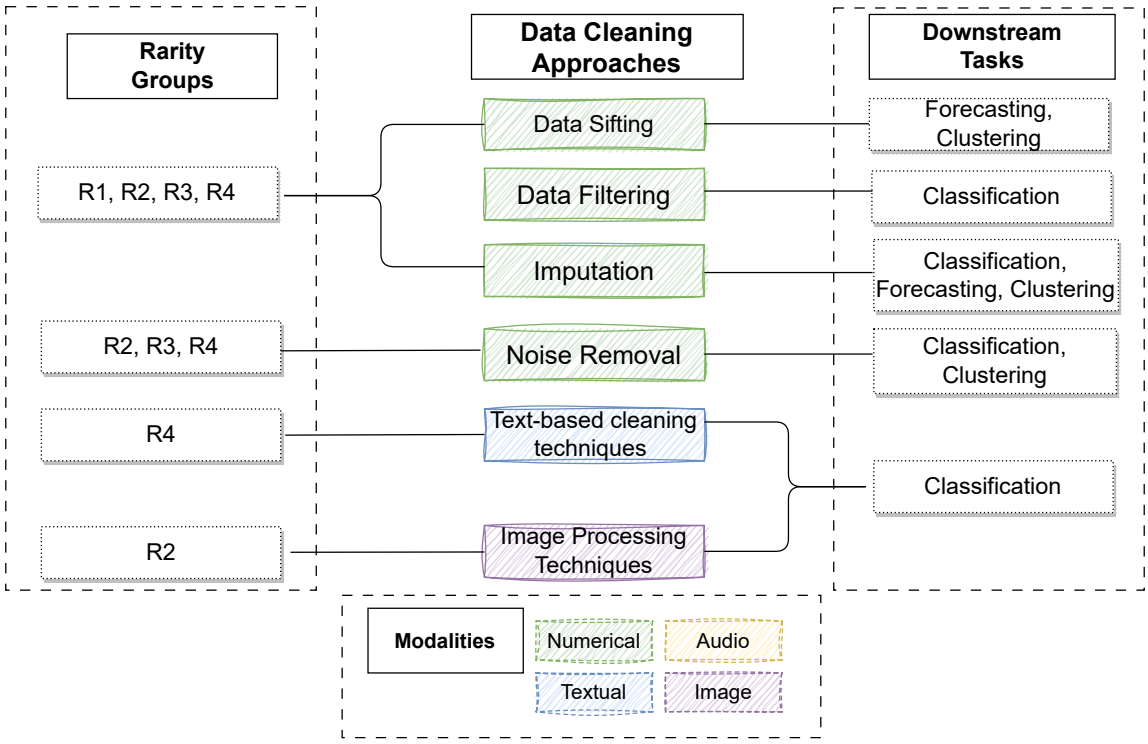

Fig. 7. Association between data cleaning approaches, data modalities, rarity groups and downstream tasks
* Coloring of data cleaning approaches corresponds to the data modalities

B.2 Analyzing feature selection approaches with data modalities, rarity groups, and downstream tasks

Figure 8 illustrates the intricate association between feature selection methods, modality variations, and rarity groups within the context of rare event prediction, as analyzed across the reviewed papers. Regarding numerical-based feature selection, feature importance, and intrinsic-based methods are independent of the rarity groups. Correlation-based feature selection, wrapper-based, filter-based, and intrinsic methods have been used with data belonging to extremely-rare and very-rare groups. TF-IDF and MFCC-based feature extraction has been used with frequently-rare data. Decision tree-based intrinsic methods were utilized with very-rare image datasets. Likewise, in data cleaning, most data processing methods support classification tasks; some have been used in clustering, regression, simulation, and forecasting research.

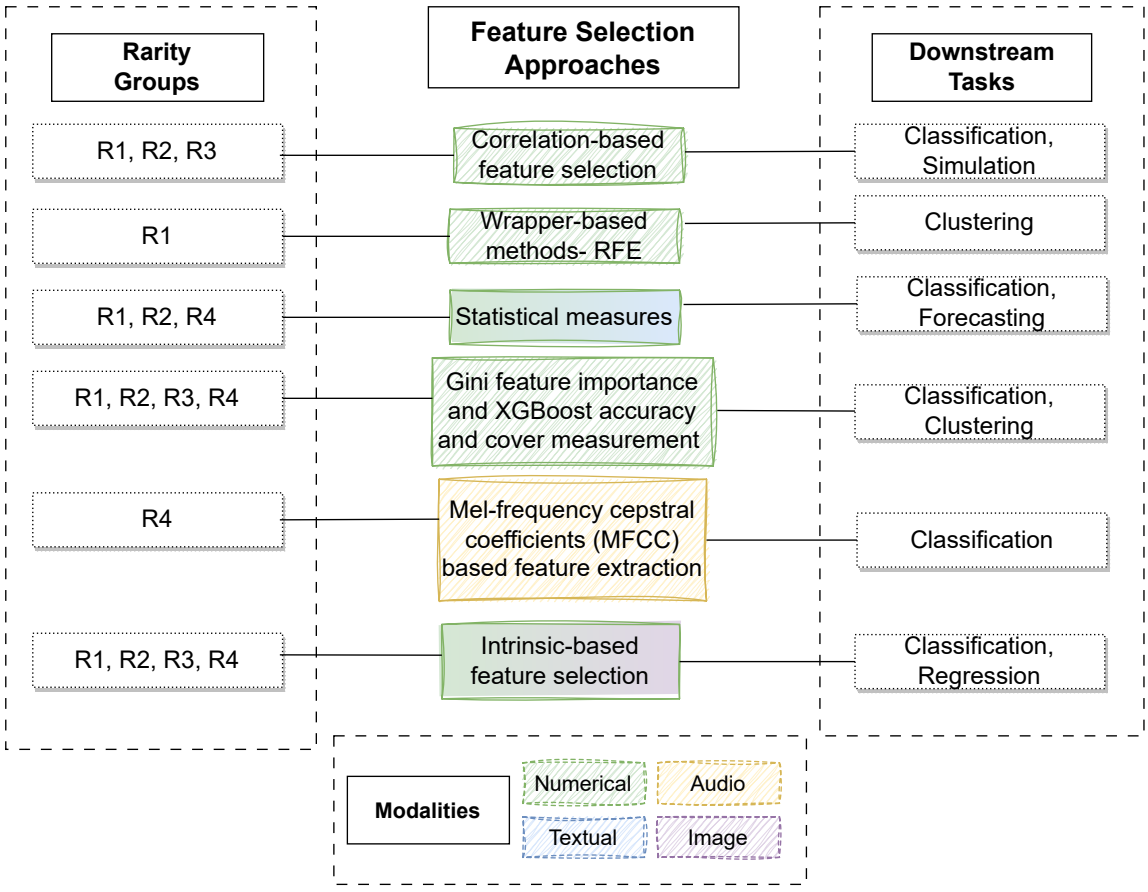

Fig. 8. Association between feature selection approaches, data modalities, rarity groups and downstream tasks
* Coloring of feature selection approaches corresponds to the data modalities

B.3 Analyzing data sampling approaches with data modalities, rarity groups, and downstream tasks

Investigating sampling approaches, we analyzed 116 rare event prediction-related papers based on sampling approach, modality, and rarity group and observed their interrelationships as depicted in Figure 9. It's seen that basic data sampling techniques are independent of rarity groups. They are used for numeric, image, and audio data and have been utilized in studies focusing on downstream tasks such as clustering, classification, and forecasting. The advanced sampling technique, SMOTE, is independent of rarity groups and is used in classification-based studies and numerical data. Most of the advanced sampling techniques, like SMUTE, ENN, NCL, NM, OSS, CBO, and time series subsampling, have

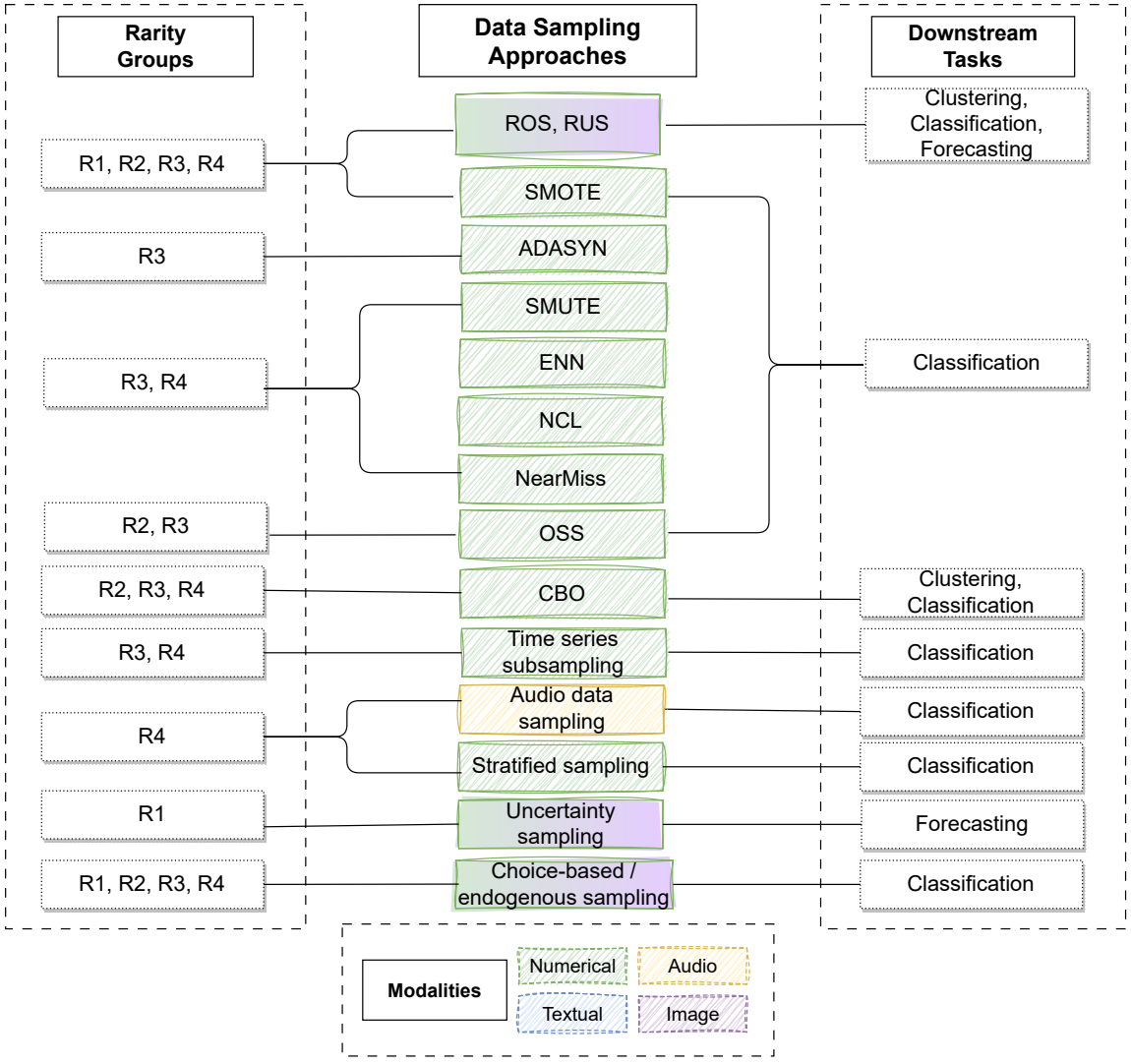

Fig. 9. Association between sampling approaches, data modalities, rarity groups and downstream tasks
* Coloring of sampling approaches corresponds to the data modalities

been widely used in rare event prediction research with numeric datasets of varying rarity levels and are mostly used in classification-based studies. Uncertainty sampling and choice-based or endogenous sampling methods have been employed for regression and forecasting tasks across different rarity groups.

B.4 Analyzing feature engineering approaches with data modalities, rarity groups, and downstream tasks

Figure 10 presents a comprehensive overview of the association between feature engineering techniques, data modality, and rarity groups in the context of rare event prediction, as examined from the reviewed papers. It is observed that classification has been the primary focus of the majority of research. Standardization, normalization, and dimensionality reduction have been applied to numerical and audio data, whereas data augmentation has been used on numerical and image data. Discretization and encoding have been used with numerical data, and in addition to classification, these techniques focus on clustering tasks. It is noted that none of the feature engineering techniques are rarity-independent; hence, each of the techniques seems to perform well with specific rarity groups.

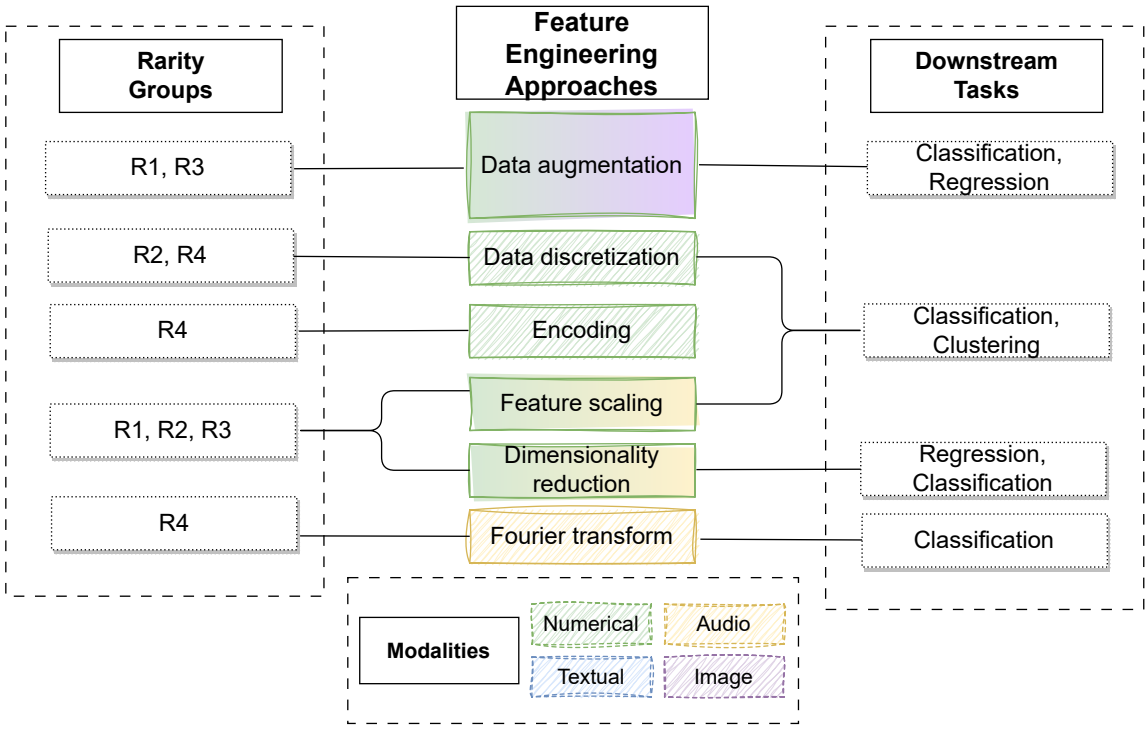

Fig. 10. Association between feature engineering approaches, data modalities, rarity groups and downstream tasks
* Coloring of feature engineering approaches corresponds to the data modalities

B.5 Analysis of data processing approaches with their primary categories, rarity groups, data modalities, and subsequent tasks: Table 6

Table 6. Data processing approaches vs. rarity groups, modality, and downstream tasks. *

Data processing approach	Papers	Rarity group	Downstream tasks	Modality	Dataset Type
1. Data Cleaning					
Data sifting	[60, 88, 117, 134]	R1, R2, R3, R4	FT, CL, CF	N	RE, DE
Data filtering	[97, 134, 201, 202]	R1, R2, R3, R4	CF	N	RE
Imputation	[38, 40, 60, 162, 170], [82, 147, 147, 157, 157, 200]	R1, R2, R3, R4	CL, CF, FT	N	RE, DE, SIY
Noise removal	[45, 100, 120, 147]]	R2, R3, R4	CL, CF	N	RE, DE
Textual summary generation	[197, 211]	R4	CF	TX	RE
Text conversions	[197]	R4	CF	TX	RE
Stemming and lemmatization	[197]	R4	CF	TX	RE
Image processing techniques	[113]	R2	CF	I	RE
2. Feature Selection					
Correlation-based feature selection	[67, 120, 146]	R1, R1, R3	CF, SM	N	RE, DE, SIY
Wrapper-based methods	[165]	R4	CL	N	RE
Statistical measures	[40, 41, 51, 82, 132, 203], [197, 211]	R1, R2, R4	CF, FT	N, TX	RE, SIY
Gini feature importance and XGBoost accuracy and cover measurement	[59, 82, 96, 147, 147]	R1, R2, R3, R4	CL, CF	N	RE
MFCC-based feature extraction	[36, 173]	R4	CF	A	RE
Intrinsic-based feature selection	[113, 120, 162]	R1, R2, R3, R4	CF, RG	N, I	RE, DE
3. Sampling					
Basic sampling i) ROS & RUS	[145, 160, 197, 198, 211], [40, 59, 89, 100, 145, 191]	R1, R2, R3, R4	CF, FT, CL	N, I, TX	RE, DE
Advanced sampling					
i) SMOTE	[42, 59, 67, 121, 170, 197, 211]	R1, R2, R3, R4	CF	N, TX	RE, DE
ii) ADASYN	[45]	R3	CF	N	RE, DE
iii) SMUTE	[120]	R3, R4	CF	N	RE, DE, SIY
iv) ENN	[45, 120, 183]	R3, R4	CF	N	RE, DE, SIY
v) NCL	[45, 120]	R3, R4	CF	N	RE, DE, SIY
vi) NearMiss	[45, 120]	R3, R4	CF	N	RE, DE, SIY
vii) OSS	[59, 113, 114, 170]	R2, R3	CF	N, I	RE, DE
viii) CBO	[104, 170]	R2, R3, R4	CL, CF	N	DE
ix) Time series subsampling	[64, 78, 123]	R3, R4	CF	N	RE
x) Stratified sampling	[123]	R4	CF	N	RE
xi) Data framing	[36]	R4	CF	A	RE
xii) Uncertainty sampling	[152]	R1	FT, SM	N	SIY
xiii) Choice-based / endogenous sampling	[127, 186]	R1,R2,R3, R4	CF	N, I	RE
4. Feature Engineering					
Data augmentation	[45, 73, 89, 150, 160, 161, 205]	R1, R3	CF, RG, CL	N, I	RE
Data discretization	[49, 144, 147, 170]	R2, R4	CL, CF	N	RE, DE
Encoding	[123]	R4	CL, CF	N	RE
Feature scaling	[36, 119, 123, 132, 133, 159, 202]	R1, R2, R4	CL, CF	N, A	RE, DE
Dimensionality reduction	[36, 51, 73, 132, 165, 203], [36, 41, 147, 162, 202]	R1, R2, R4	RG, CF, FT	N,A	RE, DE
Fourier transform	[36]	R4	CL	A	RE

* N-Numeric, TX-Textual, I-Image, A-Audio, T-Time series, FT-Forecasting, CL-Clustering, CF-Classification, RG-Regression, RE-Naturally rare, DE-Derived, SIY-Simulated/Synthetic

C ALGORITHMIC APPROACHES

C.1 Algorithmic indicators: Table 2

Explanations of the considered algorithmic indicators are as follows, which we derived from the comprehensive analysis of the papers reviewed in this study.

(1) Training time: Indicates the time required to train the model.
(2) Memory usage: Reflects the memory required for training and storing the model.
(3) Model size: Represents the model's size in terms of parameters.
(4) Feature importance: Refers to the ability of the algorithm to provide insights into the importance of different features.
(5) Model explainability: Indicates the ease of understanding the model's outcomes.
(6) Generalization: Reflects the model's ability to perform well on unseen or test data.
(7) Model complexity: Describes the complexity of the model in terms of its structure or mathematical formulation.
(8) Labeled data: Indicates that the algorithms require labeled data for training, which is typical for supervised and semi-supervised learning tasks.
(9) Unlabeled data: Leverage unlabeled data in unsupervised learning scenarios.
(10) Performance on large data: Suggests that the algorithms perform well on large datasets.
(11) Ability to handle noise: Can handle noise effectively, making them robust to noisy data.
(12) Interpretability: Enabling a better understanding of the model's decision process.

C.2 Algorithmic approaches vs. rarity groups, modality, downstream tasks, and data processing tasks: Table 7

Table 7. Algorithmic approaches vs. rarity groups, modality, downstream tasks, and data processing tasks.*

Algo. Group	Sub Algo. Group	Algo. Approach	Papers	Rarity Group	Downstream Tasks	Modality	Dataset Type	Data Processing Tasks
Supervised Classification/ Regression Methods	Threshold methods	Logistic regression	[146, 186, 211], [41, 73, 127]	R4, R1	CF, FT	I,TX, N	RE, DE, SIY	SL, FE, DC, FS
		Bayes & Neural networks	[36, 46, 120, 157], [51, 119, 150, 205]	R1,R2,R3,R4	CF	N, A, I	SI, RE, SIY	SL, FE, DC, FS
		Autoencoders	[51, 133, 202], [97, 159]	R1	CF	N	RE	SL, FE, DC, FS
	Tree-based classification methods	Random Forest	[46, 63, 82], [73, 96, 157]	R1, R2	CF	N	RE	FE, DC, FS
		Boosted Classification Trees(XGBoost, Adaboost)	[45, 73, 96], [157, 160]	R1, R2, R3	CF	N	RE	SL, FE, DC, FS
	Cost-sensitive learning	Decision tree based cost-sensitive learning, Weighting based cost-sensitive learning	[59, 95, 120, 211]	R1,R2,R3,R4	CF	N, TX	RE, SIY, DE	SL, DC, FS
	Non-parametric classification algorithms	k-nearest neighbors (k-NN)	[157]	R1, R2	CF	N	RE	FE, DC
	Kernal-based Methods	SVM	[40, 41, 100, 211], [122, 129, 157, 181]	R1, R4, R2, R3	CF, FT	N, TX	RE, SIY	SL, DC, FS
	Inference/ Rule-Based Methods	Inference methods, More appropriate inductive bias, Two phase rule induction, Utilizing knowledge and human interactions, Association rule mining	[60, 144, 200], [113, 128, 194], [105, 110, 205], [49, 60, 150]	R1, R2, R3	CF, FT	N, I	RE, DE	SL, DC, FE
Semi-Supervised & Unsupervised Methods	Random forest, PAM, K-means, Hierarchical, K-Nearest Neighbor, BIRCH, K-Medoids		[89, 147, 198], [134, 145, 165, 202]	R1, R2, R3, R4	CL	N, I	RE, DE	SL, FE, DC, FS
Statistical/ Time series Modeling	Gumbel copula function		[206]	R1	CF	N	RE	FE, DC, FS
	ARIMA		[41]	R1	CF, FT	N	DE	FE, DC, FS
	VAR		[97]	R1, R2, R3, R4	CF	N	RE	FE, DC, FS
Meta-Heuristic Optimization	Particle swarm optimization, Bat algorithm, Genetic algorithms, Evolutionary ensemble algorithms		[112, 121, 195]	R1, R2, R3, R4	CF	N	RE , DE	SL
Advanced Learning Methods	Attention-based mechanisms		[115, 123, 162, 203]	R1, R2, R3, R4	CF	N	RE	SL, FE, DC, FS
	Markov methods	Extensible markov models, Monte carlo methods	[28, 66, 134], [46, 67, 122, 146]	R1, R2, R3, R4	CF, FT, RG	N	RE, SIY, DE	SL, FE, DC, FS
	Active and Meta learning		[70, 117, 152], [88, 173]	R1	FT, SM, CF	N, I	SIY, RE	SL, FE, DC, FS

*N-Numeric, TX-Textual, I-Image, A-Audio, T-Time series, FT-Forecasting, CL-Clustering, CF-Classification, RG-Regression, RE-Naturally rare, DE-Derived, SIY-Simulated/Synthetic, SL-Sampling, FE-Feature engineering, DC-Data cleaning, FS-Feature selection

D EVALUATION APPROACHES: TABLE 8, 9,10

Table 8. General evaluation methods in rare event evaluations.*

Evaluation methodology	Sub methods	Papers	Rarity group	Algorithmic approach	Modality	Type of the dataset
Cross validation methods	K-Fold Cross-Validation	[40, 59, 63, 73], [42, 82, 113, 197]	R1, R2	SVM, Cost-sensitive learning, RF, LR, XGBoost, IBL	N, I, TX	DE, RE
	LOOCV , Stratified K-Fold	[96, 113, 119, 211]	R1, R2, R4	Cost-sensitive learning, IBL, LR, SVM, RF, XGBoost, CNN, VAR, k-NN	I, N, TX	RE
Holdout evaluation (Train-test-validation splitting)	Random Split	[63, 96, 123, 160]	R1, R4	RF, XGBoost, Attention-based, Adaboost	N	RE
	Time-Based Split	[60, 63, 67]	R1,R3,R4	Bayesian methods, RF, Monte Carlo methods	N	DE, RE
Cost and error analysis, Baseline comparison		[36, 51, 60, 67]	R1, R2, R3, R4	SVM, KNN, XGB, MLP, RF, LR	N, A, TX, I	RE, DE, SIY
Training time evaluation		[51, 119]	R1	CNN-Autoencoders, CNN, VAR, k-nearest neighbors	N	RE
Ablation studies		[89]	R1, R4	K-means, One-class learning	I, N	DE

*N-Numeric, TX-Textual, I-Image, A-Audio, T-Time series, RE-Naturally rare, DE-Derived, SIY-Simulated/Synthetic

Table 9. Rare event-based evaluation methods.*

Evaluation methodology	Papers	Rarity group	Algorithmic approach	Modality	Type of the dataset
Ahead-of-time prediction evaluation	[51, 67, 195, 202]	R1, R3	CNN-Autoencoders, CNN, Genetic algorithms	N	RE, DE
Cost-benefit analysis	[119, 144]	R1	k-nearest neighbors, Bayes and neural networks, Bayesian methods	N	RE
Root cause analysis	[119]	R2	k-nearest neighbors, Bayes and neural networks	N	RE

*N-Numeric, TX-Textual, I-Image, A-Audio, T-Time series, RE-Naturally rare, DE-Derived, SIY-Simulated/Synthetic

D.1 Analyzing evaluation approaches vs. rarity groups, modality, type of data and algorithmic approaches

Based on our literature review, the analysis of general evaluation methods and rare event-specific evaluation methods, along with their respective components, is summarized in Tables 8 and 9. The analysis highlighted that standard techniques like cross-validation have predominantly been employed in datasets encompassing numerical, textual, and image data, particularly in extremely-rare and very-rare categories. Only limited research has been dedicated to rare event-specific evaluation methods applied to numerical data. In classification and forecasting, widely used performance metrics such as accuracy, precision, recall, AUC, and ROC have been employed for evaluation, regardless of the data modalities. However, evaluation has predominantly been conducted on numerical data for other downstream tasks like clustering, regression, and simulation. Moreover, our analysis reveals the limited availability of rare event-specific evaluation techniques and highlights the need for conceptualized evaluation methodologies explicitly designed for rare event prediction. Such tailored approaches would be essential to address the unique challenges of rare events.

Table 10. Performance metrics used in Rare event evaluations.*

Downstream task	Performance metric	Papers	Rarity group	Algorithmic approach	Modality	Dataset type
Downstream task: Classification and Forecasting	Accuracy	[36, 51, 63], [67, 157, 197]	R4, R1,R2	Bayes & Neural networks, RF, Monte Carlo methods	A, N, TX	RE, DE
	Confusion matrix, Sensitivity, Specificity, FPR, FNR, Precision, F1 score	[36, 49, 51], [59, 60, 63], [67, 97, 157], [115, 195], [197, 211]	R1, R3,R4, R2	Bayes & Neural networks, Association rule mining, Cost-sensitive learning, Bayesian methods, RF, Monte Carlo methods, VAR, LSTM-autoencoder, Genetic algos	A, N, TX	RE, DE
	Geometric Mean (G-Mean)	[49, 67, 113]	R1, R3,R4, R2	Association rule mining, Monte Carlo methods, Attention-based mechanisms	N	DE, RE
	Cohen's Kappa Index	[100, 121, 205]	R1, R2, R3	SVM, Particle swarm optimization, Bat algorithm, CNN	N, I	Real
	Balanced Error Rate	[100]		SVM	N	
	Matthews's correlation coefficient	[120]	R3	Cost-sensitive learning Bayes & Neural networks	N	RE
	PR curve, ROC, AUC and AUPRC	[36, 41, 59], [42, 63, 197]	R1, R4, R2	Bayes & Neural networks, ARIMA, SVM, Cost-sensitive learning, RF	A, N, TX	DE, RE
	Cost matrix	[73]	R2	RF, LR, XGB	N	RE
	Top-decile lift	[42]		Hybrid	N	RE
	Reconstruction error	[97, 202]	R1	VAR, LSTM-autoencoder, autoencoders, hierarchical clustering	N	RE
	Moran index	[206]	R1	Statistical Modeling	N	RE
Downstream task: Clustering	Hopkins statistics test	[147]	R2	Random forest clustering, PAM	N	RE
	Silhouette coefficient	[147]	R3	Random forest clustering, PAM	N	RE
	Elbow method	[145]	R1, R2	K-Medoids clustering	N	RE
	True Skill Statistic	[40]	R1, R2	Kernel-based methods - SVM	N	DE
	Sum of Squared Errors	[165]	R4	K-nearest neighbur	N	RE
Downstream task: Regression	Mean Absolute Error and Root Mean Squared Error, Mean absolute percentage error	[162, 203]	R1, R3, R4	Attention-based mechanisms	N	RE
Downstream task: Simulation	Probability evaluation of rare events	[46, 70, 146]	R1,R2,R3,R4	Monte carlo methods, Bayes and NN, RF, LR	N	SIY
	Rare event detection metrics	[46, 146]	R1,R2,R3,R4	Monte carlo methods, Bayes and NN, RF, LR	N	SIY
	Statistical Robustness	[46]	R1,R2,R3,R4	Monte carlo methods, Bayes and NN, RF	N	SIY
	Confidence intervals	[146]	R1,R2,R3,R4	LR, Monte carlo	N	SIY

*N-Numeric, TX-Textual, I-Image, A-Audio, T-Time series, RE-Naturally rare, DE-Derived, SIY-Simulated/Synthetic

E VISION FORWARD

As the methods, data, and applications for predicting rare events continue to advance, we can see how it will affect tasks beyond prediction. We envision that it will have a significant impact on three specific areas: improving our understanding of causality, using process knowledge workflows to predict rare events, and implementing automated planning strategies to mitigate them effectively. This is illustrated in Figure 11.

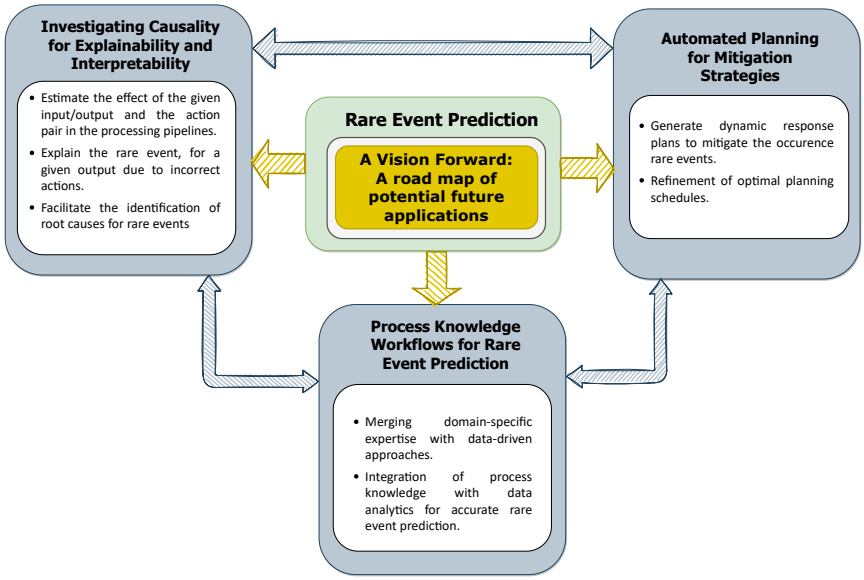

Fig. 11. The road map of potential future applications in rare event prediction

(1) Investigating causality for explainability and interpretability
An integral component of our strategy entails delving deeper into causality, which will aid us in delivering more lucid explanations. This requires us to gauge the influence of individual input-output associations and actions. In doing so, we can deduce the causes of infrequent occurrences and determine which actions by sub-systems, robots, or simulators triggered them. By comprehending causality in this manner, we can uncover the underlying reasons behind rare events and plan mitigation procedures accordingly.

(2) Process knowledge workflows for rare event prediction
Another direction for rare-event prediction entails the integration of process knowledge workflows into the prediction of rare events. This methodology enhances rare event prediction by integrating domain-specific procedural knowledge with data-driven approaches. This integration has the potential to reveal nuanced abnormalities and disparities that may not be readily apparent, hence leading to more accurate and efficient rare-event prediction.

(3) Automated planning for mitigation strategies
Lastly, the road map toward advanced rare-event prediction applications envisions the integration of automated planning techniques. Once rare events are detected and understood, the next step involves devising effective mitigation strategies. Automated planning aims to generate dynamic response plans that can mitigate the impact

of rare events. These strategies can encompass various scenarios, offering a comprehensive and adaptable framework for handling unexpected, uncertain occurrences.

In essence, a forward-thinking approach in predicting rare events relies on three crucial components: exploring causes, following knowledge workflows, and implementing automated planning. For instance, predicting rare events in manufacturing assembly pipelines, such as unexpected machinery failures or missing parts, is significant for reducing time and labor costs and improving work processes. Predicting such events involves understanding causality to identify root causes, using process knowledge workflows for accurate prediction, and implementing automated planning strategies to mitigate their impact effectively.

A key insight that we foresee here is that, these components are intricately connected, with each one capable of informing and enhancing the others. For instance, investigating causes can help detect abnormal occurrences and irregularities in knowledge workflows, while insights gained from knowledge workflows can improve automated planning techniques to prevent rare events. By identifying the causes of rare events through exploration, mitigation strategies can be developed in automated planning, and optimized strategies can be used to manipulate variables and gain a better understanding of the causal relationships between factors and rare events. Ultimately, this collaborative approach generates a more comprehensive and holistic framework for predicting rare events across various domains, as information and insights are exchanged between the three elements.

F ACRONYMS: TABLE 11

Table 11. List of acronyms

Acronym	Description	Acronym	Description
N	Numeric data	LIDAR	Light detection and ranging
TX	Textual data	LASA	Look-alike-sound-alike
I	Image data	FFT	Fast fourier transform
A	Audio data	GAN	Generative adversarial networks
T	Time series	CGAN	Conditional generative adversarial networks
CF	Classification	WGAN	Wasserstein agenerative adversarial networks
CL	Clustering	MPM	Mineral prospectivity mapping
FT	Forecasting	PCA	Principal component analysis
RG	Regression	LR	Logistic regression
SM	Simulation	NB	Naive bayes
RE	Naturally rare event datasets	NN	Neural networks
DE	Derived datasets	CNN	Convolutional neural networks
SIY	Simulated/Synthetic datasets	MLP	Multi-layer perceptron
LSTM	Long short-term memory		
UCI	University of California Irvine	RF	Random forest
KEEL	Knowledge extraction based on evolutionary learning	k-NN	K-nearest neighbors
API	Application programming interface	RIPPER	Repeated incremental pruning to produce error reduction
DC	Data cleaning	SVM	Support vector machines
FS	Feature selection	L.SVM	Support vector machine with linear kernel
SL	Sampling	R.SVM	Support vector machine with radial kernel
FE	Feature engineering	GSVM	Granular support vector machines
ML	Machine learning	RE-WKLR	Rare event weighted kernel logistic regression
R1	Extremely-rare category	MSB	Maximum specificity bias
R2	Very-rare category	IBL	Instance-based learning
R3	Moderately-rare category	1-NN	1-Nearest neighbor
R4	Frequently-rare category	PAM	Partition around medoids
CoR	Curse of rarity	CLARA	Clustering large applications
SFA	Signal fragment assembler	COG	Classification using lOcal clusterinG
VAE	Variational autoencoder	BIRCH	Balanced iterative reducing and clustering using hierarchies
DP	Data picker		
QC	Quality classifier	ARIMA	Autoregressive integrated moving average
MOA	Massive online analysis	VAR	Vector autoregression
WEKA	Waikato environment for knowledge analysis	GRU	Gated recurrent unit
PHQ	Patient health questionnaire	EMM	Extensible markov models
HIV	Human immunodeficiency virus	DNO	Deep neural operators
WWD	Wrong-way driving	FNACC	Faulty-normal accuracy
APS	Air pressure system	RFNACC	Real faulty-normal accuracy
SVD	Singular value decomposition	TP	True positives
MICE	Multiple imputation by chained equation	FN	False negatives
ANOVA	Analysis of variance	TN	True negatives
TL	Tomek links	FP	False positives
ENN	Edited nearest neighbors	TNR	True negative rate
RFE	Recursive feature elimination	FPR	False positive rate
HMM	Hidden markov model	FNR	False negative rate
DWT	Discrete wavelet transform	G-Mean	Geometric Mean
mRMR	Minimum redundancy maximum relevance	BER	Balanced error rate
TF-IDF	Term frequency-inverse document frequency	MCC	Matthews's correlation coefficient
XGBoost	eXtreme Gradient Boosting	PR	Precision-Recall
MFCC	Mel-frequency cepstral coefficients	AUPRC	Area under precision-recall curve
MDI	Mean decrease in impurity	TDL	Top-decile lift
ROS	Random minority oversampling	TSS	True skill statistic
RUS	Random majority undersampling	SSE	Sum of squared errors
SMOTE	Synthetic minority oversampling technique	MAE	Mean absolute error
ADASYN	Adaptive synthetic sampling technique	RMSE	Root mean squared error
SMUTE	Similarity majority under-sampling technique	MAPE	Mean absolute percentage error
NCL	Neighborhood cleaning rule	MAD	Mean absolute deviation
NM	NearMiss	MAPD	Mean absolute percentage deviation
NM2	NearMiss-2	AUC-ROC	Area under the receiver operating characteristic curve
OSS	One-sided selection		
CBO	Cluster-based oversampling	UQ	Uncertainty quantification
		F2F	Factory to Factory