DOI: 10.6977/IJoSI.202312_7(8).0003
H. Chaurasiya etc./Int. J. Systematic Innovation, 7(8), 24-36(2023)

Investigating feature extraction techniques for imbalanced time-series data

Harshita Chaurasiya*[1], Dr. Anand Kumar Pandey[2]
[1] Department of CSA, ITM University, Gwalior, MP, INDIA
[2] Department of CSA, ITM University, Gwalior, MP, INDIA
* Corresponding author E-mail: Harshitachaurasiya27@gmail.com

(Received 28 December 2022; Final version received 17 October 2023; Accepted 01 December 2023)

Abstract

High-class data imbalance is usually present in many applications, such as fraud detection and cancer diagnosis, hence effective classification with time-series data is an essential topic of study. Furthermore, excessively imbalanced data presents a challenge, since most learners will be biased toward the majority group, and in extreme circumstances, will overlook the minority group completely. Over the previous two decades, fundamental methodologies have been used to study class imbalance in depth. Despite recent breakthroughs in addressing data imbalance with feature extraction and its growing popularity, there is relatively little empirical work in the domain of feature extraction with time-series-based class imbalance. Following record-breaking performance outcomes in various complicated domains, researchers are now looking into the usage of feature extraction approaches for issues with significant degrees of class imbalance. To better understand the effectiveness of feature extraction when applied to class-imbalanced data, available research on class imbalance, feature extraction, and fundamental approaches like SMOTE, Resampling, and others are examined. This study explores the specifics of each study's execution and experimental outcomes, as well as provides more insight into its advantages and limitations. We discovered that there is relatively limited study in this field. Several classic approaches for class imbalance, such as data sampling and SMOTE, work with feature extraction, but more sophisticated methods that take the use of minority class feature learning abilities have potential applications. The survey continues with a discussion that identifies numerous gaps in time-series data based on class-imbalanced data to improve future studies.

Keywords: Big Data, Time Series, Machine Learning, Feature Extraction.

1. Introduction

In any domain, 'time' is the most important concept. We use the time component to plot our revenue figures, income, bottom line, and economic expansion, and even predict outcomes and estimates (Fulcher and Jones, 2017). Time series data, often referred to as time-stamped data, is a representation of the data elements that are classified in time sequence. The data which has been timestamped was collected at different points in time. Such time stamps are often made up of many data measured within the same sources over a length of time which are used to monitor variability. Because time is a component of everything observable, time series data may be found everywhere. Time series data has long been related to financial applications. Time series data is becoming more prevalent due to the increasing instrumentation of our environment (He et al., 2015). Time series analysis (TSA) is essential in understanding how variables change over time and can be applied in various sectors such as finance, retailing, academics, and meteorology (Joo, & Jeong 2019; Yang et al., 2021). TSA is often used to study non-stationary data and is useful for cluster analysis, classification, fault diagnosis, and prediction in data analysis, analytical thinking, and deep learning, as presented in Figure 1:

- TSA is used for segmentation, classifying, intrusion detection, and making predictions in

information retrieval, pattern classification, and deep learning.

- In signal analysis, industrial engineering, and information science, time series is used for signal classification and estimate.
- In statistical, inferential statistics, quantification economics, earthquake engineering, meteorological department, and geology, TSA is used for predicting.

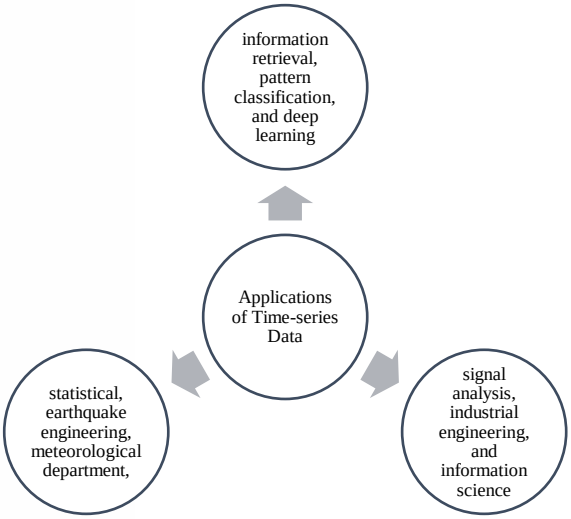

Fig.1. Application Areas of Time-Series Data

As a result, this might be a complicated issue to comprehend. When dealing with time-sensitive information, we must consider a great deal of detail in time series data (Li et al., 2018; Singh & Yassine, 2018; Shih et al., 2018; Cao et al., 2019). Existing time series forecasting methods are unquestionably effective in most circumstances, but they do have certain drawbacks. When provided merely the time component and the target variable, data scientists struggle to map their data. TSA often needs a wide range of data points in terms to maintain consistency and reliability. A big data set ensures that the sample group is representative, facilitating analytics that effectively reduces data redundancy. It also guarantees that any trends or patterns observed are not outliers and that seasonal volatility is considered (Shahriar et al., 2014; Wu & Liu 2018). This is when time series feature extraction engineering comes into play. This can turn a strong time series model into a great forecasting model.

To summarize, time-series data can be useful for classification in various domains, while imbalanced data can occur frequently in many situations. The presence of class imbalance should be considered when using time-series data for classification. Different techniques can be employed to address the class imbalance, such as resampling, cost-sensitive learning, and ensemble methods, depending on the nature of the data and the goals of the classification task.

Therefore, the purpose of this research is to meticulously examine and analyze feature extraction strategies applicable to imbalanced time-series data. Imbalances in classification data are pervasive and critical in fields like fraud detection, healthcare, and finance, often leading to models overlooking rare, yet important, events. Given that rare events, albeit infrequent, can have significant implications, the study seeks to illuminate effective methodologies to enhance the accuracy and reliability of classification models dealing with imbalanced datasets. Through a comparative analysis of various techniques, including PCA, SVD, Autoencoders, and ICA, the research aims to offer valuable insights and a foundation for the development or enhancement of algorithms that can adeptly navigate through the challenges posed by imbalanced time-series data, thereby supporting improved outcomes in detection and prediction tasks across various domains. The study also intends to identify gaps in the current literature and suggest directions for future research to advance understanding and solutions for class imbalances in time-series data.

2. Data Imbalance

A balanced dataset has an equal distribution of the target class, while an imbalanced dataset has an unequal distribution of observations where one class label has a large number of observations and the other has a small number, as in Fig 2. There are several cases in which the positive classes appear less often, such as illness diagnosis (Manogaran et al., 2019), fraud detection (Zhou, et al., 2021), computer security (Zhang et al., 2019), and picture recognition, which naturally produces skewed data distributions. Intrinsic imbalance is generated by naturally occurring data levels, particularly ones seen in clinical diagnosis where the proportion of persons is normal. Extrinsic imbalance, on the other hand, is caused by elements outside of the control of the researcher, such as collecting or storage processes (Huang et al., 2021). Supervised learning requires a training dataset with labelled samples for classification problems.

Class imbalance occurs when one class has significantly less data than another class in a binary classification problem (Krawczyk B 2016; Sun. 2011). (Often, the minority class (positive class) is the focus of attention in such cases, such as in diagnostic

imaging for disease recognition when the number of patients is limited. The majority of healthy individuals are known as negative samples in this case. Learning from these skewed datasets may be challenging, particularly when dealing with large amounts of data, and non-traditional machine learning approaches are sometimes necessary to produce good results. Because imbalanced data appears in many real-world applications, a detailed grasp of the class imbalance issue and the solutions available to remedy it is essential. This is seen in Fig. 3. The acquired information becomes meaningless and worthless if data mining algorithms are unable to categorize minority cases such as medical diagnoses of disease or abnormal products of inspection data.

This issue has recently been identified in a significant variety of real-world contexts (Zhang et al., 2022). Since the positive class has a larger probability value, learners are most prone to overclassify that when there's a class imbalance in data sets. As a result, patients from the negative class are more likely to be misidentified than cases from the positive class (Narwane & Sawarkar 2022).

Negative consequences arise from imbalanced datasets, making it challenging to accurately predict class labels. Traditional assessment criteria like accuracy can be misleading, as a naïve learner who always predicts the minority class can achieve high accuracy on a dataset with a small number of minority class samples. To overcome these issues, various classic machine-learning algorithms have been developed over the years.

To address class imbalance in machine learning, there are three types of methodologies: data-level strategies, algorithm-level methods, and hybrid approaches. Data-level strategies aim to reduce class imbalance through various data sampling approaches. Algorithm-level methods adjust the basic learners or their outcomes to reduce bias toward the dominant group, often using a weight or cost schema (Leevy et al., 2018).

Hybrid approaches (Bedi & Jindal 2021) intelligently combine data-level and algorithm-level methods. The main challenge with predicting from imbalanced datasets is accuracy, as classifiers may become biased towards the majority class. The confusion matrix displays how well the model classifies target classes, and it is used to calculate the model's accuracy in such cases.

Table 1 depicts the degree of disparity between the majority and minority classes. Imbalanced data sets are present in a broad range of classification issues that we meet in everyday banking, such as churn prediction and fraud detection. Quite often, we are confronted with severe circumstances in which the minority class fraction ranges between 0.1 and 0.2 percent. Fields such as fraud detection and anomaly detection are in critical condition. When it comes to medical diagnostics, signal errors are in a medium stage, which is straightforward to deal with.

Table 1. Degree of Imbalance

Degree of Imbalance	Proportion of the majority class	The proportion of minority class
Mild	80-60 %	20-40 %
Medium	99- 80%	1-20 %
Extreme	More than 99%	Less than 1 %

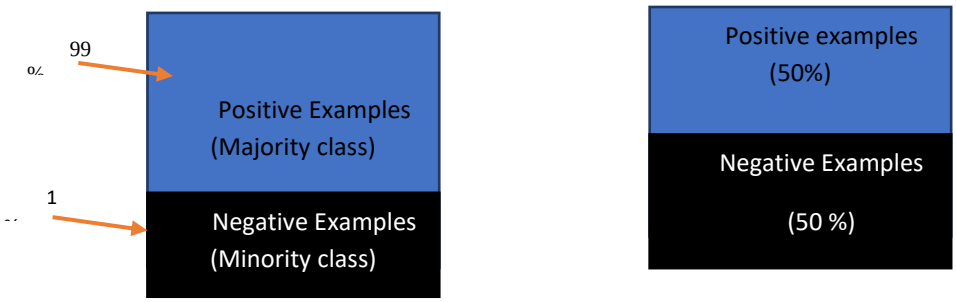

Fig. 2. Imbalanced and balanced dataset

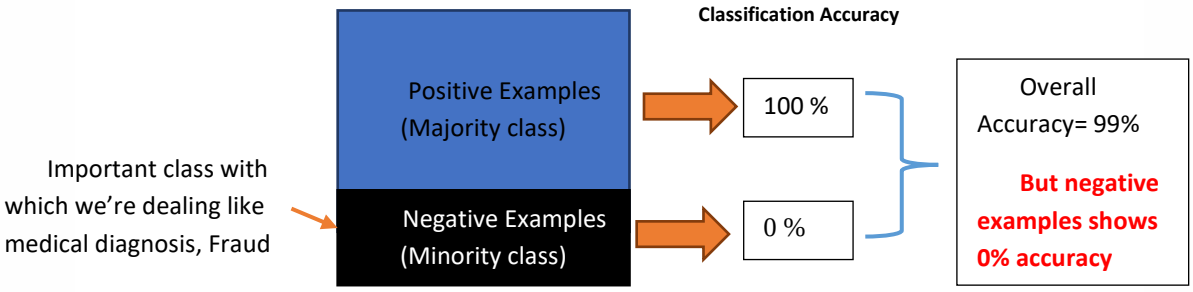

Fig. 3. Example of the class imbalance problem

3. Basic categories and challenges in data imbalance

In uncommon situations like fraud detection or illness prediction, it is crucial to accurately identify minority groups. The model should not be biased towards detecting just the dominant class but also give the minority class similar weight or relevance. Various strategies can be used to address this issue, and there is no one-size-fits-all approach to dealing with data imbalance. Each strategy performs effectively in different situations and has its own set of drawbacks.

3.1 Resampling (Oversampling and Undersampling)

The method described involves sampling a dataset to address class imbalance. Oversampling can be used to increase the number of instances in the minority class, while under-sampling can decrease the number of instances in the majority class. This can help create a balanced dataset where the classifier can treat both classes equally (Malhotra & Jain 2022; Mohammed et al., 2020). Figure 4 may provide additional illustration.

Rathpisey and Adji (2019) used resampling techniques to address the class imbalance in a hate-speech dataset, resulting in improved accuracy and effectiveness of SVM, Logistic Regression, and Naïve-Bayes models. Logistic Regression with Random Oversampling (ROS) had the highest F-1 Score of 95%. In (Lee and Kim, 2020), oversampling was used to address the imbalance in nuclear receptor profiles for deep learning predictions, resulting in a sensitivity and specificity of 71.4% and 78.7% and an accuracy rate of 82.9%, with an ROC-AUC of 0.822 using simple resampling techniques.

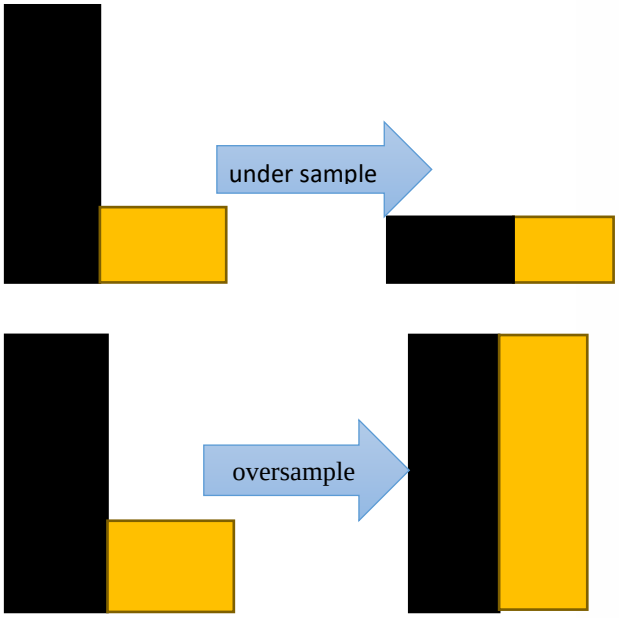

Fig.4. Under Sampling and Over-Sampling

3.2 SMOTE

SMOTE is a method of oversampling the minority class that generates new instances by using k closest neighbors to construct synthetic instances in feature space. This helps to avoid adding duplicate minority class entries to a model. Figure 5 illustrates how SMOTE generates new instances from existing data by selecting random nearest neighbors of minority class instances and creating synthetic instances in between them (Maldonado et al., 2022).

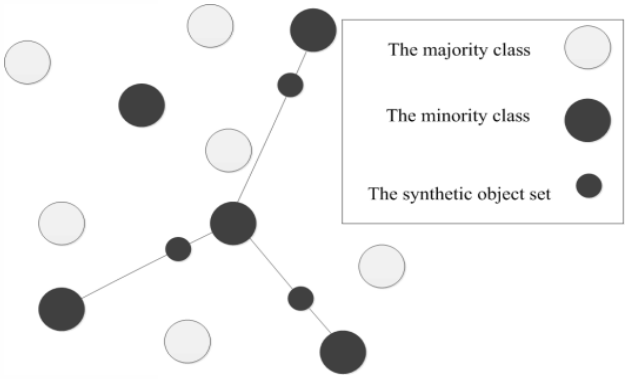

Fig.5. Basics of SMOTE Algorithm (Maldonado et al., 2022)

Rupapara et al., (2021) proposed an ensemble technique called RVVC, which combines logistic regression and support vector classifiers under soft voting rules, to detect hazardous remarks on social media networks. The performance of RVVC is evaluated on both imbalanced and balanced datasets using various evaluation metrics. The synthetic minority oversampling method (SMOTE) is used to achieve data balance on the imbalanced dataset. When TF-IDF features are employed with the SMOTE balanced dataset, RVVC outperforms all other individual models and achieves an accuracy of 0.97.

Satriaji and Kusumaningrum (2018) used the SMOTE approach with NB, SVM, and LR classifiers to balance the class distribution in a dataset. The study found that using SMOTE improved system effectiveness, particularly in imbalanced datasets, resulting in a 12% performance increase. The g-Mean score was 81.68%, with TP at 79.89% and TF-IDF at 79.31%. Among the classifiers, LR achieved the highest average score of 81.55%, followed by SVM at 81.55% and NB at 77.68%.

Pandey et al. (2019) presented an 11-layer deep-CNN mode employing SMOTE for categorizing the arrhythmias data into five classifications, per the ANSI–AAMI guidelines. The suggested novel approach has the major advantage of reducing the number of classifiers and eliminating the need to detect and divide QRS complexities. The experimental outcomes reveal that the developed Classification algorithm has better results in terms of accuracy, recall, F-1 score, and overall accuracy when compared to previous work in the research. These findings also show that the 70:30 train-test data set has the greatest performance accuracy of 98.30%.

3.3 Balanced bagging classifier

To address the issue of imbalanced datasets, a Balanced Bagging Classifier has been introduced. This classifier is similar to a traditional classifier but includes a step to balance the training set using a specific sampler during the fit. The "sampling strategy" and "replacement" parameters are used to determine the type of resampling required and whether or not to use sampling with replacement. By balancing the training set, the Balanced Bagging Classifier can help prevent the model from favoring the majority class due to its larger volume.

Ning et al. (2022), a new approach called balancing evolution semi-stacking (BESS) is proposed for sickness detection using partially labeled imbalance (PLI) data. The strategy addresses the issue of class imbalance by utilizing unsupervised learning through the BESS co-training methodology. The approach combines the information and classification diversity obtained through BESS to improve the effectiveness of the stacked ensemble. The approach is evaluated using PLI tongue image data, and the results show that BESS outperforms other state-of-the-art methods in identifying diabetic diabetes, chronic kidney illness, prostate cancer, and persistent ulcers. The statistical analysis of the results demonstrates the superiority and effectiveness of the proposed algorithm.

Table 2. The comparison between Resampling, SMOTE, and Balanced Bagging Classifier to Deal with Data Imbalance

Technique	References	Result	Advantages	Limitations
Resampling	Rathpisey and Adji (2019)	Accuracy= 91 % F1-Score= 95 %	Increase the duration of the run Reduce the amount of training data samples when the training data set is large to help with storage issues. Outperforms under sampling	Because it repeats minority class occurrences, it raises the chance of over-fitting. It can eliminate potentially helpful information that might be beneficial in the development of rule classifiers. The sample
	Lee and Kim (2020)	Sensitivity= 71.4 % specificity= 78.7 % accuracy= 82.9 %		

		and a ROC-AUC of 0.822		picked at random under-sampling might be skewed.
SMOTE	Rupapara et al., (2021)	Accuracy= 97 %	Synthetic examples are created rather than a replication of real instances, which mitigates the issue of over-fitting induced by random oversampling. There's no loss of important data.	SMOTE doesn't explore surrounding samples from different classes when producing synthesized examples. This may lead to an increase in class overlap and the introduction of extra noise. Doesn't take into account the importance of crucial traits.
	Satriaji and Kusumaningrum (2018)	G-mean = 81.68 %		
	Pandey et al., (2019)	Accuracy= 98.3 %		
Balanced Bagging Classifier	Ning et al. (2022)	The results of the trials support the suggested system's effectiveness and efficacy.	Reduces the noise Increase minor class examples to balance	Overfitting Costly

Figure 6 shows the accuracy comparison of different techniques reviewed in the literature. The accuracy of SMOTE is maximum discussed in (Rupapara et al., 2021; Pandey et al., 2019). However, it is clear that with different classifiers used the accuracy can be varied. Rupapara et al. (2021) the accuracy is maximum i.e., 98.30%.

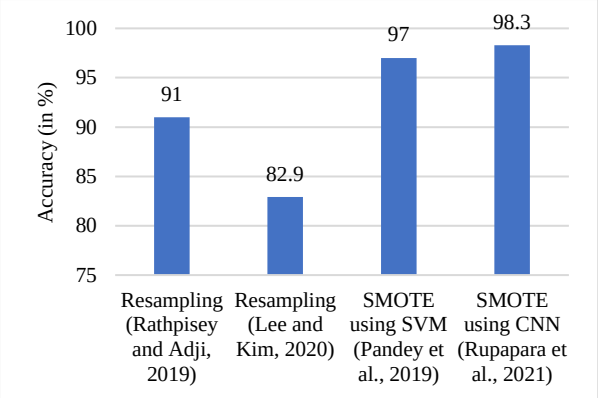

Fig. 6. The Accuracy Comparison of Different Techniques Reviewed

4. Feature extraction-based models to deal with data imbalance

Feature extraction is another technique to cope with dimensionality. Dimensionality reduction, which converts data into a low-dimensional space, is linked to feature extraction. It is important to note, however, that feature selection approaches are not the same as feature extraction techniques. Feature extraction uses functional mapping to build new features from the original data, while feature selection only returns a subset of the original features. For unstructured data such as photos, text, and audio, feature extraction techniques are more often used (Braytee et al.2016). Following are the types of feature extraction techniques discussed for our literature considering time-series data imbalance:

4.1 PCA

PCA (principal component analysis) is a method for emphasizing variance and highlighting significant patterns in a dataset. Principal Component Analysis (PCA) is a machine learning method for reducing dimensions in AI. It is a quantitative approach that transforms data of correlation qualities into a set of linear uncorrelated data via orthogonal conversion. The substantially altered qualities are the principal component. It's among the most popular data exploration and computational modelling packages. It's a technique used to extract significantly dominant points, areas, and features from a batch of data by reducing variance. When working with time-series data, it's often utilized to make data easier to understand and display.

Liu et al., (2018) concentrate on ICU death predictions, which is a frequent instance of ICU big data's secondary usage. Individual ICU death predictions are challenging for a variety of reasons, including large data, unbalanced distribution, and chronological synchronization. To discover the optimal predictor, many methods were studied, and various

AUC results were assessed in a large and important benchmark dataset. The recommended strategy surpassed the traditional machine learning approach, with SVM getting the best AUC value of 77.18%. This research establishes a framework for addressing comparable issues with large health data and aids in the promotion of healthcare services.

Abdulhammed et al. (2019) proposed using Principal Component Analysis (PCA) to reduce the dimensionality of features in the Intrusion Detection Systems (IDS) design. They showed that using low-dimensional features resulted in improved performance in terms of Detection Rate, F-Measure, False Alarm Rate, and Accuracy in binary and multi-class classification. They were able to reduce the feature dimensions of the CICIDS2017 dataset from 81 to 10 using PCA while maintaining a high accuracy of 99.6% in both multi-class and binary classification.

Hamed, et al. (2015) proposed a PCA-based technique to handle imbalanced activity data from sensor readings. A different classifier, LDA+WSVM, is utilized to address this issue. The performance of the proposed method is compared with other methods using multiple real-world datasets, and the results show that LDA+WSVM achieved a higher recognition rate. The recall, precision, F-score, and accuracy of the proposed method are reported to be 77%, 78.4%, 77.7%, and 93.5%, respectively.

4.2 Singular value decomposition

Similar to principal component analysis, SVD is a data decomposition method (PCA). It is used in signal processing and statistics for a variety of tasks, including signal feature extraction, matrix approximation, and pattern identification (Chang et al.2012).

SVD exists and is unique up to the signs for any matrix X (m by n). For the data matrix X, the singular value decomposition is:

$$X = UDV^t \quad (1)$$

Where, U, V= Left and Right Singular Vector, D= singular vectors' diagonal

The choice of a small number (k) of additional features is based on criteria for the proportion of initial data variation accounted by the new features (usually 80-90 percent). In these extra features, every one of the original features has indeed been evaluated. SVD could be used straight for feature extraction since the additional features (columns of U) integrate the old features (columns of X). Uses rank constraints on the SVD to ensure that each new primary coordinate has a minimal number of nonnegative. This will result in the extraction of original characteristics as well as their meaning (Modarresi, 2015).

To address this challenge, Chen, et al. (2008) proposed a unique paradigm termed the "Information Granulation Based Data Mining Approach." Information Granules, rather than numerical data, are used to gain knowledge in the suggested approach, which mimics the human capacity to digest information. Experiments demonstrate that strategy can considerably improve the ability to categorize data that is skewed. The proposed method's overall accuracy and G-mean are 0.973 and 0.948 respectively. Hossain & Rab, (2022) proposed a novel approach involving SVD and a modified ensemble classifier that outperformed the extreme learning machine (ELM) with macro-F1 scores of 90.78 percent. All of the deep learning techniques outperformed these benchmarks. The ensemble classifier excelled on the Reuters dataset, with an accuracy of 91.49 percent. They created four datasets with varying degrees of imbalance to experiment. A modified ensemble classifier based on the results was also given, which can classify both imbalanced and balanced data.

4.3 Autoencoder

An autoencoder, a kind of NN, could be used to generate a condensed version of raw input. Figure 7 depicts the encoding and decoding sub-models that make up an auto-encoder. The encoding block compresses the data, and the decoding block attempts to rebuild it from the encoding block's compressed form. Throughout the training, the encoding model is preserved, but the decoding model is removed. After that, the encoding could be used as a data method of preparation to extract features from the data to train a new ML model.

The suggested DlapAE method with Laplacian regularization in (Zhao et al.2020) may enhance this fault diagnostic framework's generalization performance and make it more suited for feature learning and classification of imbalanced data. Last but not least, two examples of experimental bearing systems may be used to demonstrate the efficacy of the suggested technique. In comparison to previous deep learning-based fault diagnosis approaches, the suggested fault diagnostic method can successfully execute accurate fault detection for balanced and imbalanced rotating machinery datasets. The proposed

approach is 99.4% accurate and has a standard deviation of 0.2514. The dataset has a recognition accuracy of 96.94%, while the standard deviation in the 10 trials is 1.7%.

Alhassan et al. (2019) used a prediction-based DL method to help in the assessment of individual overall mortality in healthcare. The Stacked Denoised AE was trained on an intrinsically unbalanced time-stamped database. Different computational intelligence algorithms that employ various data balancing methodologies are compared with the performance. The suggested model, which surpasses typical DL algorithms with an accuracy of 0.7713, intends to solve the issue of imbalanced data.

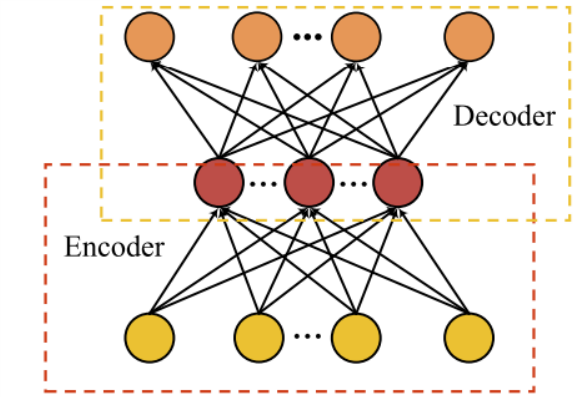

Fig. 7. Diagram of Autoencoder

Wong et al. (2020) proposed two cost-sensitive algorithms to solve the class imbalance issue: Cost-Sensitive Deep Neural Networks (CSDNN) and Cost-Sensitive Deep Neural Network Ensembles (CSDE). CSDNN is a variant of Stacked Denoising Autoencoders, while CSDE is the ensemble learning version of CSDNN. The proposed approaches are evaluated using six real-life datasets in diverse business areas. The results demonstrate that the suggested approaches work well in handling class imbalance problems and outperform all other methods tested. The performance metrics used are TPR, AUC, and G-mean, and the cost-sensitive learning algorithms AdaCost, CSDNN, and MetaCost reach 0.64, 0.58, and 0.52, respectively.

4.4 ICA

ICA (Independent Component Analysis): ICA is a continuous dimension reduction technology that takes a set of non-dependent components as input and seeks to correctly identify each one while eliminating any non-useful noise. Two input characteristics are said to be non-dependent when their linear and nonlinear dependencies are equal to zero. Independent Component Analyses are widely used in medical applications such as EEG and MRI evaluation to identify significant data from unproductive ones.

On a limited and imbalanced fMRI dataset from a word-scene memory challenge, Wang, et al. (2020) presented a three-step technique. Convolution-GRU for time series-imbalance dataset, and also Independent Component Analysis was used. The suggested technique has a 72.2 percent accuracy, which increases classification performance. The findings reveal that the suggested approaches work well in dealing with class imbalance problems and outperform all other methods tested.

Yang et al. (2021) built a novel two-channel hybrid CNN for automatic ECG recognition utilizing time series imbalanced data using ICA, with an accuracy of 0.9554 for identifying normal, CHF, and CAD individuals using leave-one-out cross-validation. Tests with multi-level noisy and unbalanced data yielded similarly excellent results. As a result, the suggested approach can identify coronary artery disease (CAD) and congestive heart failure (CHF) in clinical settings.

Table 3. The Comparison Between Different Feature Extraction Techniques

Technique	References	Result	Advantages	Disadvantages
PCA	Liu et al. (2018)	AUC performance of 0.7718 is achieved	Every new main dimension should have a restricted number of nonzero factors. Correlated Features are removed. Enhances the Algorithm Overfitting is reduced as a result of performance.	Information Loss Data standardization is a must before PCA Independent variables become less interpretable
	Abdulhammed et al., (2019)	high accuracy of 99.6%		
	Abidine et al. (2015)	Recall= 77 % F-score= 78.4 % Accuracy= 78.4 %		
SVD	Chen et al. (2008)	Accuracy= 97.33 % % G-Mean = 94.83 % %	Simplifies data removes noise may improve algorithm results	Transformed data may be difficult to understand information loss
	Hossain et al. (2022)	F-1 Score= 90.78 % Accuracy= 71.449 %		

Autoencoder	Zhao et al. (2019)	Accuracy= 99.4 %	With a non-linear activation function and several layers, it is possible to learn non-linear transformations. Instead of learning one large transformation using PCA, it is more efficient to use an autoencoder.	Insufficient training data Training the wrong use case Too lossy Misunderstanding important variables Better alternatives.
	Alhassan et al. (2018)	Accuracy= 77.13 %		
	Leung et al. (2020)	TPR= 64 % AUC= 58 % G-Mean= 52%		
ICA	Wang et al. (2020)	Accuracy= 72.2 %	By changing the input space into a maximally independent basis, information may be separated. Simple to comprehend	Problem with overfitting Costly Loss of information Require more computation as compared to PCA

Figure 8 shows the accuracy comparison of various feature extraction methods. For PCA discussed by Abdulhammed, et al., 2019), the model has the highest accuracy almost 100 % as it Removes Correlated Features and also reduces overfitting while predicting. For AE (Alhassan, et al., 2019) the accuracy is also good compared with other methods.

Figure 9 shows the comparison of SMOTE, SVD, and AE considering G-Mean as a performance parameter. G-mean is an important parameter while dealing with class imbalance problems. The basic formula of G-mean is given below:

$$\text{G-Mean} = \sqrt{Specitivity * Recall}$$

As it has Specificity and recall the chances of considering the majority class is very low. Therefore, giving better performance. The SVD discussed by Chen et al., (2008) has the highest G-Mean nearly about 97.33%. For SMOTE discussed by Satriaji, et al. (2018), G-mean is 91.68 % and for AE it is minimum i.e., 52%.

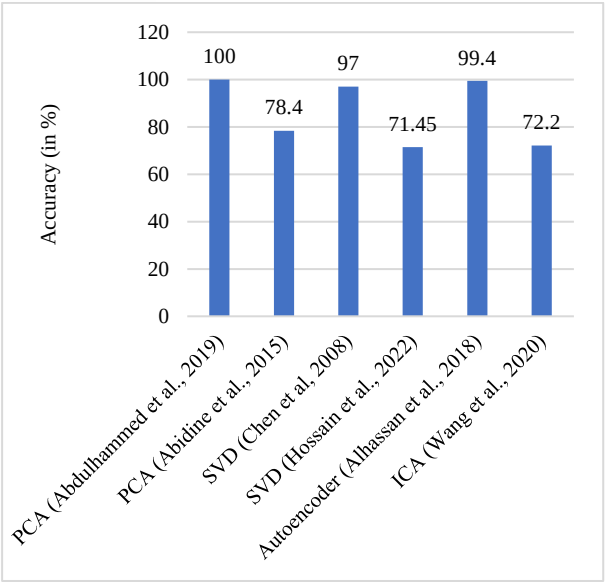

Fig.8. Accuracy Comparison of Feature Extraction based Techniques

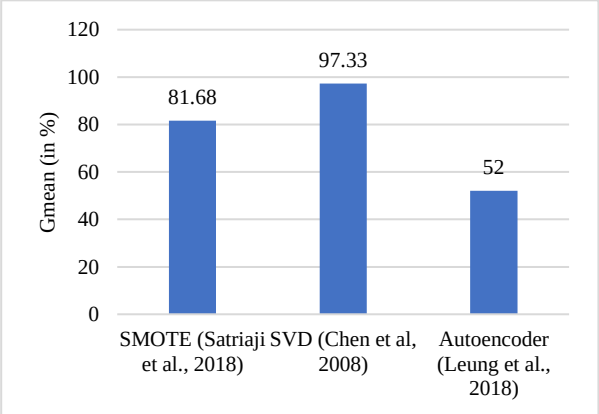

Fig- 9 G-mean Parameter comparison between different methods

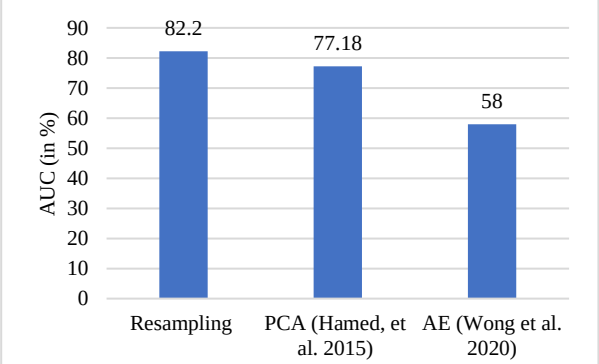

Fig- 10 AUC Parameter comparison between different methods

Figure 10 shows the comparison of Resampling, PCA, and AE using AUC as a performance parameter. It is clear from the graph that AUC is the maximum for a resampling method which is a very basic method to solve the data imbalance problem. The AE discussed by Wong, et al. (2020) has a minimum AUC of nearly about 58%.

Figure 11 shows the comparison of Resampling, PCA, and SVD using the F-1 score as a performance parameter. It is clear from the graph that for resampling

it is maximum. For PCA (Hamed, et al. 2015) it is 74.4 % and for SVD it is 70.78 %.

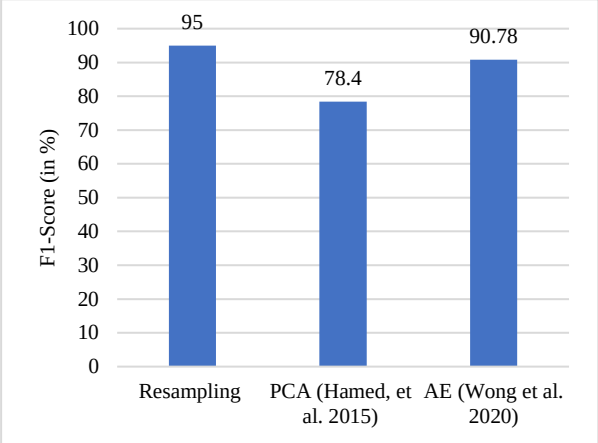

Fig- 11 F-1 Score Parameter comparison between different methods

5 Challenges and future research directions

The article suggests that there is a need for more sophisticated feature extraction approaches to address the class imbalance in time-series data. The focus would be on developing new techniques or adapting existing ones to handle imbalanced data. Ensemble methods and deep learning techniques are potential research directions that could improve classification performance on imbalanced time-series datasets. By improving feature extraction, researchers can improve the accuracy and reliability of classification models for imbalanced time-series data, which would be useful in areas such as fraud detection and disease diagnosis.

6 Conclusion

The task of detecting rare events is important in many domains, such as healthcare, signal processing, and finance. However, identifying these events can be challenging due to their infrequency and casual nature. Misclassifying rare events can lead to significant financial losses or even health risks. The detection process is also hindered by the problem of imbalanced data classification, where rare events are of greater interest but occur less frequently, making it difficult for classification algorithms to learn from them.

To address this challenge, the paper examines various feature extraction strategies for imbalanced time-series data. The goal is to develop effective approaches for feature extraction that can improve the accuracy and reliability of classification models for imbalanced data.

The paper compares the performance of different methods, including resampling, PCA, SVD, and autoencoder. Based on the given results, it can be concluded that the resampling method performs better than the other two feature extraction methods (PCA and AE) in terms of AUC and F-1 score for imbalanced time-series data classification. Resampling is a simple and effective technique to address imbalanced data, but it may result in overfitting or underfitting, depending on the type of resampling method used. PCA and SVD can help reduce the dimensionality of the data and extract relevant features, but they may not be effective in capturing complex relationships or patterns in the data.

Future research can focus on improving security management via time–series imbalanced learning. Because of the rapid rise of social media, internet security has gotten a lot of attention in recent years.

References

Abdulhammed, R., Faezipour, M., Musafer, H., & Abuzneid, A. (2019). Efficient network intrusion detection using PCA-based dimensionality reduction of features. *2019 International Symposium on Networks, Computers and Communications, ISNCC 2019*. https://doi.org/10.1109/ISNCC.2019.8909140

Alhassan, Z., Budgen, D., Alshammari, R., Daghstani, T., McGough, A. S., & Al Moubayed, N. (2019). Stacked Denoising Autoencoders for Mortality Risk Prediction Using Imbalanced Clinical Data. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 541–546. https://doi.org/10.1109/ICMLA.2018.00087

Bedi, P., Gupta, N., & Jindal, V. (2021). I-SiamIDS: an improved Siam-IDS for handling class imbalance in network-based intrusion detection systems. *Applied Intelligence*, *51*(2), 1133–1151. https://doi.org/10.1007/S10489-020-01886-Y/FIGURES/14

Braytee, A., Liu, W., & Kennedy, P. (2016). A cost-sensitive learning strategy for feature extraction from imbalanced data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9949 LNCS*, 78–86. https://doi.org/10.1007/978-3-319-46675-0_9

Chang, C. D., Wang, C. C., & Jiang, B. C. (2012). Singular value decomposition based feature

extraction technique for physiological signal analysis. *Journal of Medical Systems*, *36*(3), 1769–1777. https://doi.org/10.1007/S10916-010-9636-3

Chen, M. C., Chen, L. S., Hsu, C. C., & Zeng, W. R. (2008). An information granulation based data mining approach for classifying imbalanced data. *Information Sciences*, *178*(16), 3214–3227. https://doi.org/10.1016/J.INS.2008.03.018

Fulcher, B. D., & Jones, N. S. (2017). hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. *Cell Systems*, *5*(5), 527-531.e3. https://doi.org/10.1016/J.CELS.2017.10.001

Hamed, M. ', Abidine, B., Fergani, B., & Oualkadi, A. El. (2015). News Schemes for Activity Recognition Systems Using PCA-WSVM, ICA-WSVM, and LDA-WSVM. *Information 2015, Vol. 6, Pages 505-521*, *6*(3), 505–521. https://doi.org/10.3390/INFO6030505

He, G., Duan, Y., Peng, R., Jing, X., Qian, T., & Wang, L. (2015). Early classification on multivariate time series. *Neurocomputing*, *149*(PB), 777–787. https://doi.org/10.1016/J.NEUCOM.2014.07.056

Hossain, T., Mauni, H. Z., & Rab, R. (2022). Reducing the Effect of Imbalance in Text Classification Using SVD and GloVe with Ensemble and Deep Learning. *COMPUTING AND INFORMATICS*, *41*(1), 98-115–198–115. https://doi.org/10.31577/CAI_2022_1_98

Huang, C. Y., Dai, H. L., Huang, C. Y., & Dai, H. L. (2021). Learning from class-imbalanced data: review of data driven methods and algorithm driven methods. *Data Science in Finance and Economics 2021 1:21*, *1*(1), 21–36. https://doi.org/10.3934/DSFE.2021002

Jian Cao, Zhi Li, Jian Li, (2019). Financial time series forecasting model based on CEEMDAN and LSTM, *Physica A: Statistical Mechanics and its Applications*, 519, 2019, 127-139. https://doi.org/10.1016/j.physa.2018.11.061

Joo, Y., & Jeong, J. (2019). Under Sampling Adaboosting Shapelet Transformation for Time Series Feature Extraction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11624 LNCS*, 69–80. https://doi.org/10.1007/978-3-030-24311-1_5/COVER

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/S13748-016-0094-0/TABLES/1

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*(1), 1–30. https://doi.org/10.1186/S40537-018-0151-6/TABLES/5

Lee, Y. O., & Kim, Y. J. (2020). The Effect of Resampling on Data-imbalanced Conditions for Prediction towards Nuclear Receptor Profiling Using Deep Learning. *Molecular Informatics*, *39*(8), 1900131. https://doi.org/10.1002/MINF.201900131

Li, L., Wu, Y., Ou, Y., Li, Q., Zhou, Y., & Chen, D. (2018). Research on machine learning algorithms and feature extraction for time series. *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, *2017-October*, 1–5. https://doi.org/10.1109/PIMRC.2017.8292668

Liu, J., Chen, X. X., Fang, L., Li, J. X., Yang, T., Zhan, Q., Tong, K., & Fang, Z. (2018). Mortality prediction based on imbalanced high-dimensional ICU big data. *Computers in Industry*, *98*, 218–225. https://doi.org/10.1016/J.COMPIND.2018.01.017

Maldonado, S., Vairetti, C., Fernandez, A., & Herrera, F. (2022). FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification. *Pattern Recognition*, *124*, 108511. https://doi.org/10.1016/J.PATCOG.2021.108511

Malhotra, R., & Jain, J. (2022). Predicting defects in imbalanced data using resampling methods: an empirical investigation. *PeerJ. Computer Science*, *8*. https://doi.org/10.7717/PEERJ-CS.573

Manogaran, G., Shakeel, P. M., Hassanein, A. S., Malarvizhi Kumar, P., & Chandra Babu, G. (2019). Machine Learning Approach-Based Gamma Distribution for Brain Tumor Detection and Data Sample Imbalance Analysis. *IEEE Access*, *7*, 12–19. https://doi.org/10.1109/ACCESS.2018.2878276

Modarresi, K. (2015). Unsupervised Feature Extraction Using Singular Value Decomposition. *Procedia Computer Science*, *51*(1), 2417–2425. https://doi.org/10.1016/J.PROCS.2015.05.424

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication*

Systems, ICICS 2020, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Narwane, S. V., & Sawarkar, S. D. (2022). *Comparative Analysis of Machine Learning Algorithms for Imbalance Data Set Using Principle Component Analysis*. 103–115. https://doi.org/10.1007/978-981-16-9650-3_8

Ning, Z., Ye, Z., Jiang, Z., & Zhang, D. (2022). BESS: Balanced evolutionary semi-stacking for disease detection using partially labeled imbalanced data. *Information Sciences*, *594*, 233–248. https://doi.org/10.1016/J.INS.2022.02.026

Pandey, S. K., & Janghel, R. R. (2019). Automatic detection of arrhythmia from imbalanced ECG database using CNN model with SMOTE. *Australasian Physical & Engineering Sciences in Medicine*, *42*(4), 1129–1139. https://doi.org/10.1007/S13246-019-00815-9

Rathpisey, H., & Adji, T. B. (2019). Handling Imbalance Issue in Hate Speech Classification using Sampling-based Methods. *Proceeding - 2019 5th International Conference on Science in Information Technology: Embracing Industry 4.0: Towards Innovation in Cyber Physical System, ICSITech 2019*, 193–198. https://doi.org/10.1109/ICSITECH46713.2019.8987500

Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model. *IEEE Access*, *9*, 78621–78634. https://doi.org/10.1109/ACCESS.2021.3083638

Satriaji, W., & Kusumaningrum, R. (2018). Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis. *2018 2nd International Conference on Informatics and Computational Sciences, ICICoS 2018*, 99–103. https://doi.org/10.1109/ICICOS.2018.8621648

Shahriar, M. S., Rahman, A., & McCulloch, J. (2014). Predicting shellfish farm closures using time series classification for aquaculture decision support. *Computers and Electronics in Agriculture*, *102*, 85–97. https://doi.org/10.1016/J.COMPAG.2014.01.011

Singh, S., & Yassine, A. (2018). Big Data Mining of Energy Time Series for Behavioral Analytics and Energy Consumption Forecasting. *Energies 2018, Vol. 11, Page 452*, *11*(2), 452. https://doi.org/10.3390/EN11020452

Shih, S. Y., Sun, F. K., & Lee, H. yi. (2018). Temporal Pattern Attention for Multivariate Time Series Forecasting. *Machine Learning*, *108*(8–9), 1421–1441. https://doi.org/10.1007/s10994-019-05815-0

Wang, S., Duan, F., & Zhang, M. (2020). Convolution-GRU Based on Independent Component Analysis for fMRI Analysis with Small and Imbalanced Samples. *Applied Sciences 2020, Vol. 10, Page 7465*, *10*(21), 7465. https://doi.org/10.3390/APP10217465

Wong, M. L., Seng, K., & Wong, P. K. (2020). Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain. *Expert Systems with Applications*, *141*, 112918. https://doi.org/10.1016/J.ESWA.2019.112918

Wu, J., Yao, L., & Liu, B. (2018). An overview on feature-based classification algorithms for multivariate time series. *2018 3rd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2018*, 32–38. https://doi.org/10.1109/ICCCBDA.2018.8386483

Yang, J. Y., Hu, H. W., Liu, C. H., Chen, K. Y., Un, C. H., Huang, C. C., Chen, C. C., Lin, C. C. K., Chang, H., & Lin, H. M. (2021). Differencing Time Series as an Important Feature Extraction for Intradialytic Hypotension Prediction using Machine Learning. *3rd IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability, ECBIOS 2021*, 19–20. https://doi.org/10.1109/ECBIOS51820.2021.9510749

Yang, W., Si, Y., Zhang, G., Wang, D., Sun, M., Fan, W., Liu, X., & Li, L. (2021). A novel method for automated congestive heart failure and coronary artery disease recognition using THC-Net. *Information Sciences*, *568*, 427–447. https://doi.org/10.1016/J.INS.2021.04.036

Zhang, T., Chen, J., Li, F., Zhang, K., Lv, H., He, S., & Xu, E. (2022). Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions. *ISA Transactions*, *119*, 152–171. https://doi.org/10.1016/J.ISATRA.2021.02.042

Zhang, Y., Chen, X., Guo, D., Song, M., Teng, Y., & Wang, X. (2019). PCCN: Parallel Cross Convolutional Neural Network for Abnormal Network Traffic Flows Detection in Multi-Class

Imbalanced Network Traffic Flows. *IEEE Access*, *7*, 119904–119916. https://doi.org/10.1109/ACCESS.2019.2933165

Zhou, X., Hu, Y., Liang, W., Ma, J., & Jin, Q. (2021). Variational LSTM Enhanced Anomaly Detection for Industrial Big Data. *IEEE Transactions on Industrial Informatics*, *17*(5), 3469–3477. https://doi.org/10.1109/TII.2020.3022432

Zhao, X., Jia, M., & Lin, M. (2020). Deep Laplacian Auto-encoder and its application into imbalanced fault diagnosis of rotating machinery. *Measurement*, *152*, 107320. https://doi.org/10.1016/J.MEASUREMENT.2019.107320.