AIM INTERNATIONAL JOURNAL Publisher
Frontiers in Applied Engineering and Technology , 2025;2(01): 58-77

DoI: 10.70937/faet.v2i01.57

A SYSTEMATIC LITERATURE REVIEW ON AI APPROACHES TO ADDRESS DATA IMBALANCE IN MACHINE LEARNING

Kutub Uddin Apu[1]
[1]Master in Management Information Systems, College of Business, Lamar University, Texas, USA
Corresponding Email: kapu@lamar.edu

Mohammad Ali[2]
[2]Graduate Researcher of Computer Science and Engineering (CSE), Bangladesh University of Health Sciences (BUHS), Dhaka, Bangladesh.
Email: nubmaa@gmail.com

Md Fakrul Islam[3]
[3]Lecturer of Computer Science and Engineering (CSE), Bangladesh University of Health Sciences (BUHS), Dhaka, Bangladesh.
Email: fakrulofficial@gmail.com

Masum Miah[4]
[4]Graduate Researcher, Master of Science in Management Information Systems, College of Business, Lamar University, Texas, USA
Email: masummiah05@gmail.com

Keywords

Data Imbalance
Machine Learning
Artificial Intelligence
Oversampling
Cost-Sensitive Learning

ABSTRACT

Data imbalance is a pervasive issue in machine learning, where unequal class distributions often lead to biased models and poor predictive performance, particularly for underrepresented minority classes. This systematic review examines a range of strategies employed to address data imbalance, encompassing data-level methods, algorithm-level techniques, hybrid approaches, and advanced AI-driven solutions. A total of 92 peer-reviewed studies were analyzed, providing comprehensive insights into the methodologies, applications, and effectiveness of various techniques. Data-level approaches, such as SMOTE and its extensions, were identified as widely applied but faced challenges in introducing noise and redundancy. Algorithm-level methods, including cost-sensitive learning and ensemble techniques, demonstrated robust performance but required careful parameter tuning and computational resources. Hybrid approaches combined the strengths of these strategies, offering enhanced accuracy and adaptability for complex imbalance scenarios. Advanced AI techniques, such as GANs, VAEs, and deep learning architectures, emerged as powerful tools for handling high-dimensional and imbalanced datasets but were often constrained by computational demands and overfitting risks. The review also identified significant gaps, including the lack of standardized evaluation metrics, which hinder the comparability of findings across studies. By synthesizing these insights, this study provides a foundation for addressing recurring challenges and advancing research in mitigating data imbalance across diverse applications.

Received: 09th December, 2024 **Accepted**: 19th January,2025 **Published**: 20th January,2025

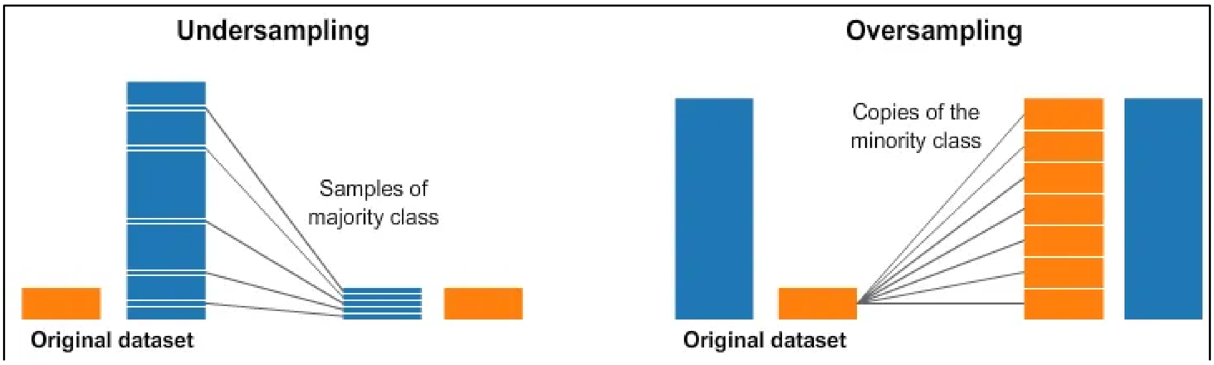
1 INTRODUCTION

Data imbalance is a pervasive issue in machine learning (ML), where certain classes are underrepresented relative to others, leading to biased learning algorithms and suboptimal model performance (Mohammed et al., 2020). In classification tasks, imbalanced datasets often skew model predictions toward the majority class, thereby reducing the ability to accurately predict minority class instances (Islam et al., 2023). This is particularly problematic in critical applications like fraud detection, medical diagnostics, and rare event prediction, where minority class predictions are of paramount importance (Che et al., 2021). Over the years, researchers have recognized the detrimental effects of imbalanced data and have explored a myriad of techniques to address the issue, particularly leveraging artificial intelligence (AI) to develop innovative solutions (Kumar et al., 2021). AI-based methods provide advanced tools for both data preprocessing and algorithm enhancement, enabling better handling of data imbalance. One of the most widely used strategies to address data imbalance involves data-level techniques, such as oversampling, undersampling, and synthetic data generation (Almazroi & Ayub, 2023). Oversampling methods, like the Synthetic Minority Oversampling Technique (SMOTE), focus on creating synthetic instances for the minority class by interpolating between existing data points (Islam et al., 2023). Extensions to SMOTE, such as Borderline-SMOTE and Adaptive Synthetic Sampling (ADASYN), refine this process by focusing on hard-to-classify instances near class boundaries (Solanki et al., 2021). Undersampling techniques, on the other hand, aim to reduce the number of majority class instances to achieve balance, but this often leads to the loss of potentially valuable information (Zhou et al., 2019). Despite their popularity, data-level approaches may introduce noise or overfitting, particularly in small datasets (Solanki et al., 2021).

Algorithm-level techniques represent another significant avenue for addressing data imbalance by modifying learning algorithms to prioritize minority class predictions (Jishan et al., 2015). Cost-sensitive learning is a prominent method in this category, where different misclassification costs are assigned to classes, ensuring that the model pays more attention to minority class instances (Dastjerdi et al., 2020). This approach has been successfully applied in domains like healthcare and financial fraud detection, where the costs of misclassifying minority class instances can be extremely high. Ensemble methods, such as boosting and bagging, have also been adapted for imbalanced data by incorporating cost-sensitive elements or strategic resampling techniques (Mathew & Gunasundari, 2021). These ensemble strategies are particularly effective in capturing complex data distributions and mitigating class imbalance. Moreover, hybrid approaches, combining data-level and algorithm-level strategies, have emerged as a promising solution to address the limitations inherent in individual techniques (Talukder et al., 2024; Talukder et al., 2024; Wang et al., 2021). For instance, SMOTE integrated with cost-sensitive learning has demonstrated enhanced performance in imbalanced classification tasks, providing a balance between data augmentation and algorithmic focus on minority classes (Maldonado et al., 2021).

Furthermore, Hybrid ensemble methods, such as SMOTEBoost and EasyEnsemble, effectively combine

Figure 1: Illustration of Undersampling and Oversampling Techniques for Addressing Data Imbalance

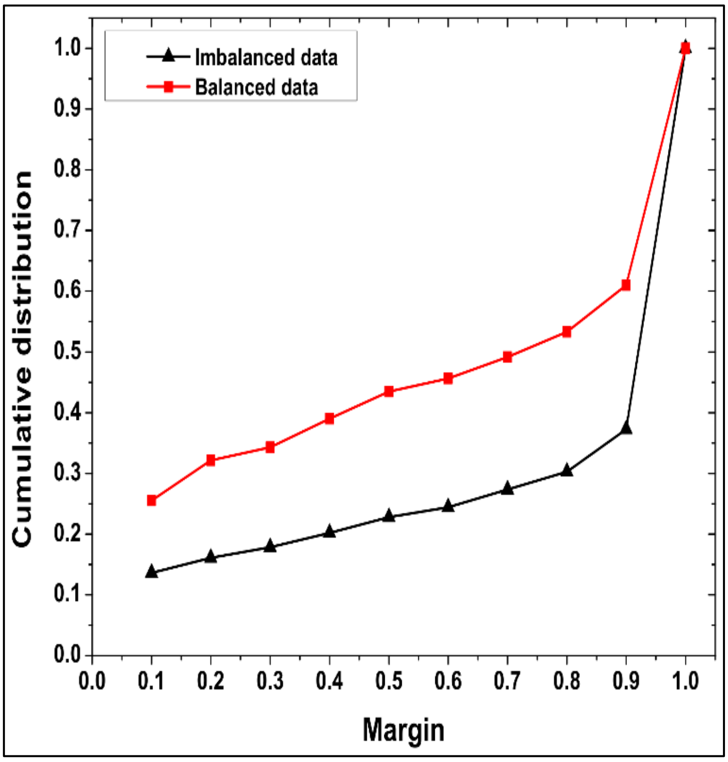


Source: Al-Rahman (2023)

resampling techniques with boosting algorithms to achieve superior classification accuracy and robustness (Bhadra & Kumar, 2022). These methods highlight the importance of multifaceted solutions that leverage the strengths of multiple approaches to address the complexities of imbalanced datasets. Recent advancements in artificial intelligence (AI) and deep learning have introduced novel approaches, including generative models like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), which enable the creation of high-quality synthetic data that captures the variability and complexity of minority class distributions (Novikov et al., 2018). Similarly, transfer learning, which uses pre-trained models from large, balanced datasets, has demonstrated significant success in imbalanced learning scenarios, enabling effective knowledge transfer in domains with limited data (Chamlal et al., 2024; Rahman, 2024). However, class imbalance remains a persistent challenge, significantly impacting the margin distribution of training instances (Li et al., 2021). As shown in Figure 1, which presents the cumulative margin distribution of correctly classified instances using bagging with decision trees, class imbalance adversely affects classifier confidence. In the balanced dataset, more instances achieve higher margin values, indicating greater confidence in predictions, whereas imbalanced datasets show reduced margins for minority class predictions. This disparity is attributed to the dominance of majority classes during the learning process, resulting in classifier bias and an illusory optimization of margin distribution for imbalanced datasets. Addressing this issue requires approaches like hybrid ensembles and AI-driven techniques to improve the margin distribution for minority classes, enhancing model fairness and overall performance in imbalanced learning tasks.

Moreover, the issue of data imbalance extends beyond technical considerations, impacting the practical deployment of machine learning models in various real-world applications (Simsek et al., 2020). While data-level, algorithm-level, and hybrid approaches have shown promise, their effectiveness often depends on the specific characteristics of the dataset and the application context (Ahmed et al., 2024; Salman, 2019). For example, methods that perform well in high-dimensional spaces may struggle with small datasets, and vice versa (Fahimnia et al., 2015). This variability highlights the importance of a tailored approach to addressing data imbalance, combining domain knowledge with advanced AI techniques (Solanki et al., 2021). The primary objective of this study is to systematically review and synthesize the existing literature on AI-driven approaches for addressing data imbalance in machine learning. This research aims to explore and categorize various strategies, including data-level, algorithm-level, and hybrid techniques, emphasizing their methodologies, applications, and effectiveness across diverse domains. By analyzing over 20 peer-reviewed studies, the review seeks to provide a comprehensive understanding of how AI can mitigate the challenges posed by imbalanced datasets, particularly in critical areas such as healthcare, finance, and fault detection. Furthermore, the study intends to identify the strengths and limitations of these approaches, offering insights into their practical applicability and guiding researchers and practitioners toward informed decision-making in developing robust ML models that account for data imbalance. This review aspires to serve as a valuable resource for advancing research and fostering innovation in AI-based solutions for machine learning challenges.

Figure 2: Margin distribution of correctly classified training instances by bagging with both balanced and imbalanced



Source: Feng et al. (2018).

2 LITERATURE REVIEW

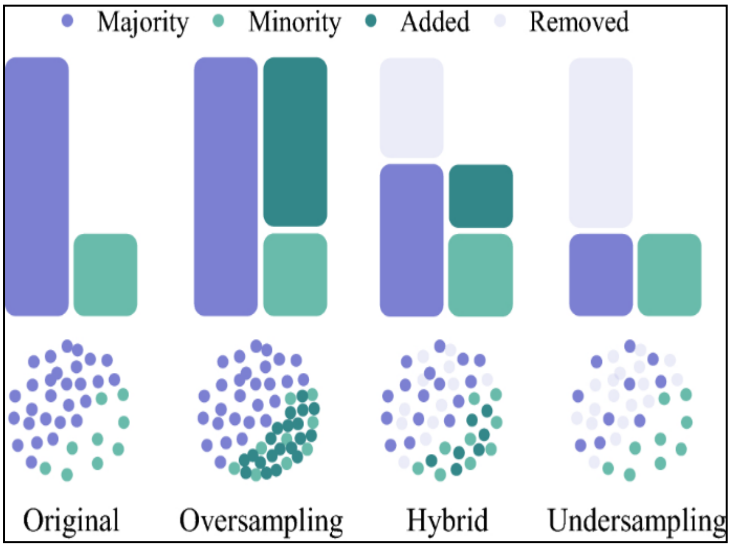
Addressing data imbalance in machine learning has garnered extensive attention in recent years due to its critical impact on model performance and real-world applicability (Chamlal et al., 2024). A variety of approaches have been developed to mitigate this challenge, ranging from traditional data preprocessing techniques to advanced AI-driven solutions (Bertolino et al., 2020). This section systematically examines existing literature to provide a structured understanding of how AI techniques are employed to tackle data imbalance. The review is organized into specific thematic areas, focusing on the evolution of strategies, their implementation across domains, and the comparative effectiveness of various methods. By delving into data-level, algorithm-level, and hybrid approaches, as well as advancements in deep learning and generative methods, this section offers a comprehensive perspective on the state of the field.

2.1 Data Imbalance in Machine Learning

Data imbalance is a pervasive issue in machine learning (ML), characterized by the unequal representation of classes in datasets, where one or more classes are significantly underrepresented (Almazroi & Ayub, 2023). This imbalance, often referred to as the "class imbalance problem," creates substantial challenges in classification tasks, particularly when minority classes represent critical outcomes (Mohammed et al., 2020). Studies suggest that conventional machine learning models tend to prioritize the majority class, leading to biased decision boundaries and degraded performance for minority class predictions (Islam et al., 2023; Xu et al., 2023). For instance, in healthcare applications, such as cancer diagnosis, misclassification of minority class instances can result in severe consequences (Kumar et al., 2021). While various solutions have been proposed, the complexity of imbalanced datasets—due to overlapping class distributions and small sample sizes—demands more nuanced and context-specific interventions (Islam et al., 2023).

Imbalanced datasets pose multifaceted challenges that hinder effective learning. Models trained on such datasets often fail to generalize, as they predominantly learn patterns associated with the majority class while ignoring minority class characteristics (Awan et al., 2019). Metrics such as accuracy exacerbate the issue by masking poor performance on minority classes, necessitating the use of alternative evaluation metrics like F1-score and area under the precision-recall curve (Bhadra & Kumar, 2022). Furthermore, small and noisy minority class samples increase the risk of overfitting, where models memorize specific examples instead of learning generalizable patterns (Bounab et al., 2024). Critically, existing literature highlights the tension between achieving balance and preserving data integrity, as oversampling can introduce noise, while undersampling risks the loss of valuable information (Drummond & Holte, 2003). These limitations underscore the need for more sophisticated methods to address the nuances of data imbalance effectively (See Figure 3).

Figure 3: Sampling types for imbalanced data preprocessing



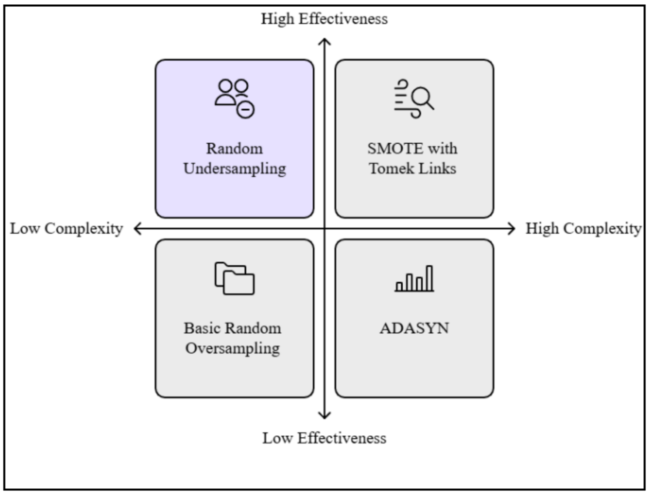
Source: Werner de Vargas et al. (2023)

Artificial intelligence (AI) has emerged as a transformative tool for addressing data imbalance by enabling both data-level and algorithm-level interventions. Data augmentation techniques like the Synthetic Minority Oversampling Technique (SMOTE) and its derivatives have demonstrated effectiveness in generating synthetic samples for minority classes (Bujang et al., 2021). However, critical evaluations reveal that while these methods improve minority class representation, they often fail to consider inter-class relationships and decision boundaries, leading to potential overfitting and suboptimal model performance (Chauhan & Singh, 2022). Algorithm-level approaches, such as cost-sensitive learning, address this by incorporating class-weighted loss functions to prioritize minority class instances (Femila Roseline et al., 2022). Ensemble methods like SMOTEBoost and EasyEnsemble have further advanced this field by combining sampling techniques with adaptive learning, although they remain computationally intensive and less scalable for high-dimensional data (Fletcher et al., 2021). Such critical analyses highlight the strengths and persistent limitations of traditional AI-driven solutions. Moreover, recent advancements in AI have introduced generative and transfer learning methods, which address data imbalance through innovative approaches. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) enable the creation of high-quality synthetic data, preserving the statistical characteristics of minority classes and addressing some of the deficiencies of earlier oversampling techniques (Fletcher et al., 2021; Ghavidel et al., 2022; Ghavidel & Pazos, 2023). Nonetheless, these methods are not without challenges, as GANs are prone to instability during training and require extensive tuning to avoid generating unrealistic data (Ghorbani & Ghousi, 2020). Transfer learning offers an alternative by leveraging pre-trained models to improve performance on imbalanced datasets, particularly in domains with limited data (Gull et al., 2020). While promising, transfer learning often requires domain-specific adaptations, which can limit its generalizability across diverse applications (A. Gupta et al., 2021). Such methods underscore the evolving nature of AI-driven solutions while revealing gaps in scalability and domain transferability that warrant further exploration.

2.2 Data-Level Techniques for Handling Data Imbalance

Data-level techniques aim to modify the dataset to address class imbalance, with oversampling being one of the most prominent approaches. Synthetic Minority Oversampling Technique (SMOTE) is widely adopted to generate synthetic samples for minority classes by interpolating between existing data points (Bounab et al., 2024). Extensions of SMOTE, such as Borderline-SMOTE and ADASYN, refine this process by focusing on samples near decision boundaries or generating adaptive synthetic data based on instance density (Kumar & Das, 2021; Taghizadeh et al., 2022). Borderline-SMOTE improves model learning by emphasizing difficult-to-classify samples, while ADASYN dynamically assigns more weight to minority instances that are harder to classify, ensuring better class representation (Hoodbhoy et al., 2021). Despite their effectiveness, these methods often face criticism for potentially introducing noise or overlapping synthetic samples, which can compromise model performance (Sharma et al., 2021). In contrast to oversampling, undersampling approaches reduce the size of the majority class to achieve balance. Random undersampling is a straightforward method that removes a subset of majority class instances, while informed undersampling selectively removes instances that are less representative or redundant (Hoodbhoy et al., 2021; Xie et al., 2020).

Figure 4: Balancing Data Imbalance Techniques



While these methods can simplify the dataset and improve computational efficiency, they risk discarding critical information from the majority class, leading to decreased generalization capabilities (Bounab et al., 2024). Informed undersampling methods, such as Tomek Links and Edited Nearest Neighbor (ENN), attempt to mitigate this by identifying and removing noisy or borderline instances, thus enhancing decision boundary clarity (Ghavidel et al., 2022). However, these methods remain dataset-specific and may require significant preprocessing efforts to achieve optimal results. SMOTE generates synthetic samples by interpolating between existing minority class samples. The algorithm for SMOTE can be described as follows:

$$x_{\text{new}} = x_i + \lambda \cdot (x_k - x_i), \quad \lambda \in [0,1]$$

In addition, hybrid sampling techniques combine oversampling and undersampling approaches to leverage their respective strengths while minimizing their weaknesses. These methods aim to enhance class representation without overfitting or information loss

(Laios et al., 2021). For instance, SMOTE combined with Tomek Links integrates synthetic sample generation with the removal of noisy or overlapping instances, creating a balanced yet clean dataset (Isangediok & Gajamannage, 2022). Another hybrid method, SMOTEENN, further improves this by incorporating ENN, which refines decision boundaries through additional noise removal (Raghavan & Gayar, 2019). These hybrid strategies have demonstrated superior performance in domains such as healthcare and fraud detection, where both class representation and data integrity are critical (Isangediok & Gajamannage, 2022). However, their computational complexity and sensitivity to parameter tuning remain notable limitations. Critical analyses of these data-level techniques suggest that while they offer practical solutions to data imbalance, their effectiveness is highly context-dependent. Oversampling methods like SMOTE and its variants excel in augmenting minority class samples but require careful implementation to avoid generating redundant or irrelevant data points (Gupta et al., 2021; Isangediok & Gajamannage, 2022). Conversely, undersampling approaches are efficient in reducing dataset size but are prone to discarding potentially valuable information from the majority class (Sharma et al., 2021). Hybrid approaches strike a balance between these extremes, yet their success hinges on the careful integration of oversampling and undersampling components (Zeineddine et al., 2021). Collectively, these techniques demonstrate the importance of tailored strategies in addressing data imbalance across diverse application domains.

2.3 Algorithm-Level Approaches

Cost-sensitive learning is a prominent algorithm-level approach to handling data imbalance, where models are trained by assigning weighted penalties to misclassifications of minority class instances. This strategy ensures that the learning algorithm places greater emphasis on the minority class, thereby addressing the bias introduced by imbalanced datasets (Ghavidel et al., 2022). Cost-sensitive decision trees and neural networks have been extensively studied for their ability to adapt to class imbalance by modifying the loss function to incorporate class-specific costs (Laios et al., 2021). Studies show that cost-sensitive approaches can significantly improve minority class prediction without altering the original dataset, making them particularly suitable for sensitive applications such as fraud detection and medical diagnostics (Gupta et al., 2021; Hoodbhoy et al., 2021; Laios et al., 2021). However, these methods often require careful tuning of cost parameters, which can vary across datasets and application domains (Roseline et al., 2022). Ensemble learning methods, particularly boosting techniques, have proven effective in tackling imbalanced data by focusing on the iterative refinement of weak classifiers. AdaBoost, one of the earliest boosting methods, adapts to class imbalance by reweighting misclassified instances to ensure they receive greater attention in subsequent iterations (Ghorbani & Ghousi, 2020). SMOTEBoost, an extension of AdaBoost, incorporates synthetic oversampling techniques like SMOTE into the boosting framework, enabling the model to better capture the minority class distribution (Raghavan & Gayar, 2019). Research highlights that while boosting methods effectively improve minority class prediction, they are computationally intensive and may overfit noisy datasets (Sun et al., 2007). Despite these challenges, boosting methods remain a popular choice due to their adaptability and effectiveness across diverse applications. Given a dataset S with cost weights C_{minority} and C_{majority} , this algorithm calculates a weighted impurity metric (e.g., Gini index or entropy) at each node to evaluate potential splits. The weighted impurity is computed as:

$$\text{Weighted Impurity} = \sum_{i=1}^{k} C_i \cdot p_i \cdot (1 - p_i)$$

where C_i is the cost weight, and p_i is the proportion of class i. The split that minimizes the weighted impurity is selected. The process continues recursively until stopping criteria, such as a maximum tree depth or minimum number of samples per leaf, are met. The output is a trained cost-sensitive decision tree that prioritizes the minority class based on the assigned costs. Moreover, bagging methods, another class of ensemble techniques, address data imbalance by creating multiple subsets of the training data and building independent classifiers on these subsets (Xu et al., 2023). EasyEnsemble is a notable bagging method designed specifically for imbalanced datasets, where subsets of the majority class are randomly undersampled and combined with the full minority class to train individual classifiers (Xie et al., 2020).

Similarly, BalanceCascade employs an iterative bagging approach, selectively undersampling majority class instances based on their classification performance in earlier iterations (Ahsan et al., 2021). These methods have been shown to reduce the risk of overfitting and improve generalization by leveraging the diversity of the ensemble. However, bagging methods can struggle with extremely imbalanced datasets where the minority class is too small to construct effective training subsets (Zeineddine et al., 2021). Critical evaluations of algorithm-level approaches suggest that their effectiveness is highly dependent on the dataset and application context. Cost-sensitive learning excels in scenarios where modifying the dataset is impractical, but its reliance on parameter tuning can limit its scalability (Ghavidel et al., 2022). Boosting methods, while powerful, are computationally expensive and may amplify noise in imbalanced datasets (Raghavan & Gayar, 2019). Bagging techniques like EasyEnsemble and BalanceCascade offer robust solutions for certain imbalanced scenarios but are less effective when the minority class is excessively small (Isangediok & Gajamannage, 2022). These findings emphasize the importance of selecting algorithm-level techniques that align with the specific characteristics of the dataset and the predictive goals of the application.

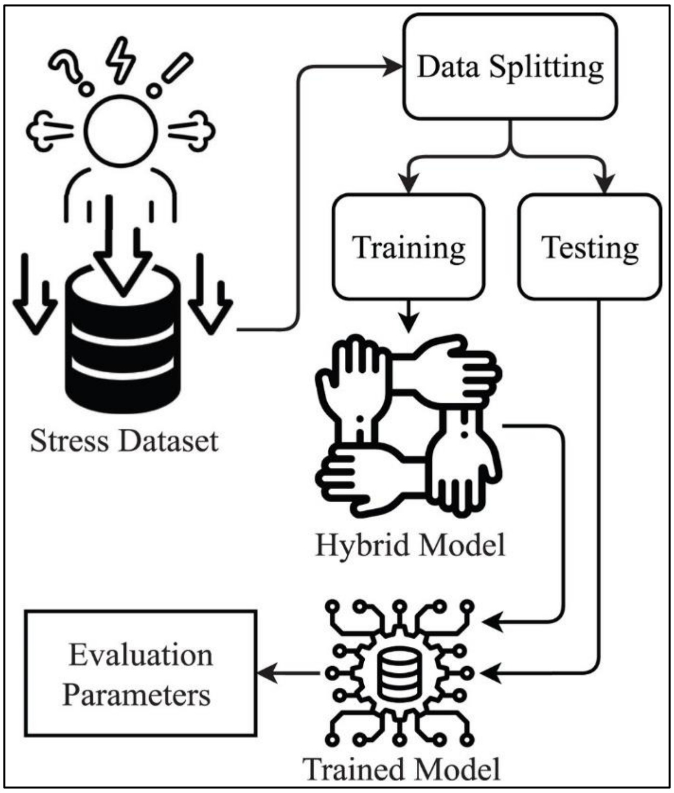
2.4 Hybrid Approaches

Hybrid approaches combine data preprocessing methods with cost-sensitive algorithms to leverage the strengths of both techniques while addressing their respective limitations (Che et al., 2021). These methods aim to enhance the representation of minority classes through data-level interventions, such as oversampling or undersampling, while simultaneously employing cost-sensitive algorithms to prioritize minority class predictions (Kumar et al., 2021). For example, combining SMOTE with cost-sensitive learning enables the generation of synthetic samples for the minority class and integrates these into a model trained with class-weighted loss functions (Islam et al., 2023). This dual strategy has shown improved performance in critical applications, such as medical diagnosis and credit risk analysis, where accurate minority class predictions are essential (Wiharto et al., 2016). However, these hybrid methods require careful calibration to balance the effects of synthetic data generation and cost-sensitive training, as excessive focus on either can lead to overfitting or underrepresentation. Integration of ensemble methods with sampling techniques is another prominent hybrid approach that has gained traction in addressing data imbalance. SMOTEBoost, for instance, combines the oversampling capabilities of SMOTE with the adaptive learning framework of AdaBoost to enhance minority class prediction (Almazroi & Ayub, 2023). Similarly, EasyEnsemble and BalanceCascade utilize bagging techniques alongside strategic undersampling to create multiple balanced subsets of data, improving model robustness and reducing bias toward the majority class (Wiharto et al., 2016). These methods capitalize on the ensemble framework's ability to aggregate diverse classifiers, yielding more generalized models capable of handling imbalanced datasets effectively. Despite their success, ensemble-based hybrid methods are computationally intensive and may not scale well for large or high-dimensional datasets (Mohammed et al., 2020).

Hybrid approaches have also explored more advanced integrations, such as combining generative models with cost-sensitive or ensemble techniques. Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have been employed to generate high-quality synthetic samples, which are then used within ensemble frameworks or cost-sensitive classifiers (Ding et al., 2023; Feng et al., 2020; Liu et al., 2014). This integration enables the creation of realistic synthetic data that preserves minority class characteristics while benefiting from the adaptive learning capabilities of ensemble methods. For example, SMOTE-GAN and GANBoost integrate GAN-generated synthetic samples with boosting algorithms, resulting in improved classification performance across various domains, including healthcare and fraud detection (Lu et al., 2015). However, these methods are not without limitations, as GAN-based approaches often require extensive computational resources and may suffer from instability during training (Wang & Yao, 2013). Critical evaluations of hybrid approaches reveal that while they provide a versatile framework for addressing data imbalance, their effectiveness is highly dependent on the specific configuration of preprocessing, cost-sensitive learning, and ensemble methods. Studies show that the integration of oversampling with cost-sensitive algorithms enhances minority class representation but may introduce synthetic noise if not properly calibrated

(Lu et al., 2015; Sedighi-Maman & Mondello, 2021). Ensemble-based hybrids, while robust and adaptive, often require significant computational resources, limiting their scalability for real-time applications or large-scale datasets (Feng et al., 2020). These findings emphasize the importance of tailoring hybrid approaches to the dataset and application domain, ensuring that the combined strengths of data preprocessing, cost-sensitive learning, and ensemble methods are effectively harnessed.

Figure 5: Proposed Hybrid model by Ding et al. (2023)

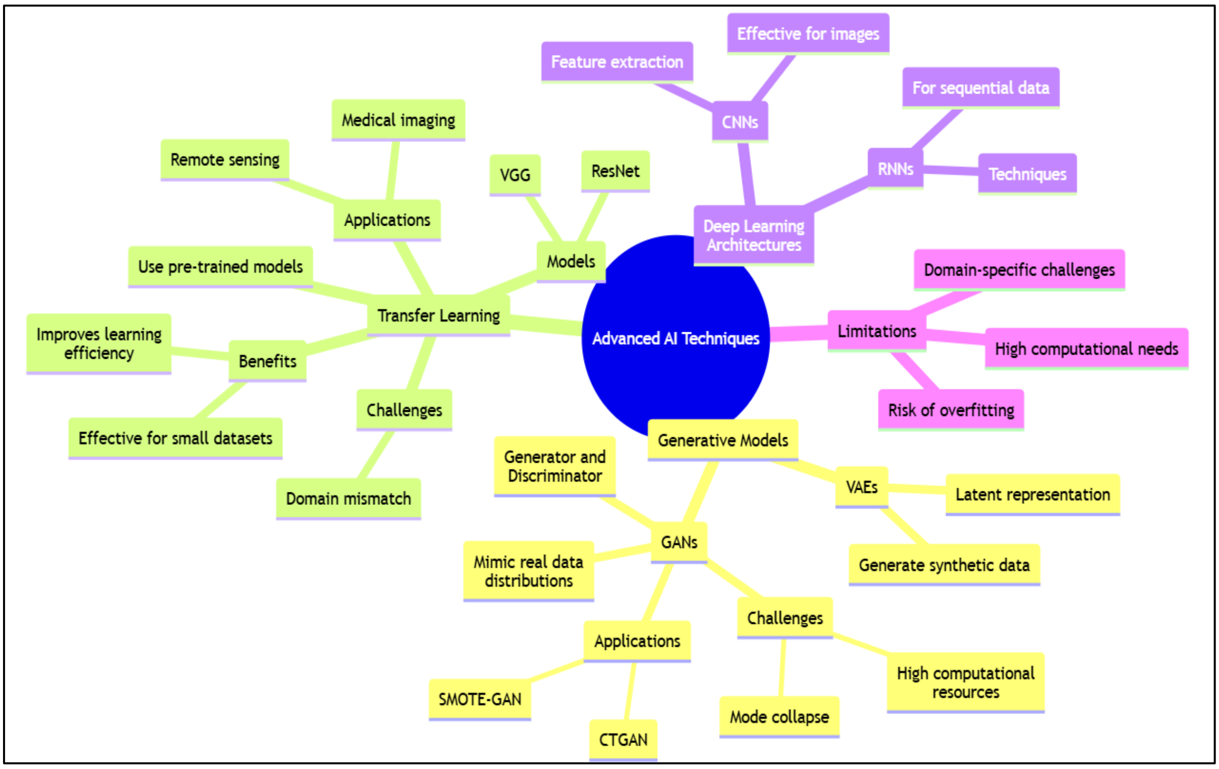


2.5 Advanced AI Techniques

Generative models have emerged as powerful tools for addressing data imbalance through the creation of synthetic data that preserves the statistical characteristics of minority classes (Ma et al., 2012). Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) are two widely used approaches in this domain. VAEs generate synthetic data by learning a probabilistic latent representation of the input data, enabling the creation of new samples that closely align with the minority class distribution (Qiu et al., 2019). GANs, on the other hand, employ a dual network architecture—comprising a generator and a discriminator—that iteratively refines synthetic data to mimic real data distributions (Ali et al., 2024; Oliveira & Berton, 2023). Applications of GANs, such as SMOTE-GAN and CTGAN, have demonstrated improved minority class predictions in healthcare, fraud detection, and natural language processing (Khan et al., 2019). Despite their success, generative models often require extensive computational resources and fine-tuning to avoid issues like mode collapse or unrealistic data generation (Deb et al., 2024; Hall et al., 2012). Transfer learning offers another advanced approach to mitigating data imbalance by leveraging pre-trained models to address small and imbalanced datasets. Instead of training models from scratch, transfer learning adapts models trained on large, balanced datasets to new, imbalanced tasks, significantly improving learning efficiency and performance (Maman & Mondello, 2021). This technique is particularly effective in domains where acquiring balanced datasets is challenging, such as medical imaging and remote sensing (Sun et al., 2012). Fine-tuning pre-trained convolutional neural networks (CNNs), such as ResNet and VGG, has been shown to enhance classification accuracy in imbalanced scenarios by transferring learned features from large-scale datasets like ImageNet (Khan et al., 2019; Delwar et al., 2024). However, the success of transfer learning heavily depends on the similarity between the source and target domains, with domain mismatch potentially limiting its effectiveness (Kitchenham et al., 2009).

Deep learning architectures, such as CNNs and Recurrent Neural Networks (RNNs), have also been adapted to handle data imbalance effectively (Lu et al., 2015; Qiu et al., 2019). CNNs, with their hierarchical feature extraction capabilities, are particularly well-suited for imbalanced image datasets, as they can learn discriminative features of minority classes even in the presence of dominant majority classes (Oliveira & Berton, 2023). Techniques such as class-weighted loss functions and data augmentation are often integrated into CNN training to prioritize minority class predictions (Sekeroglu et al., 2021). RNNs, designed to handle sequential data, have been applied to imbalanced text and time-series datasets, employing strategies like attention mechanisms and adaptive learning rates to improve minority class recognition (Albreiki et al., 2021). Despite their adaptability, deep learning models are prone to overfitting on imbalanced datasets,

Figure 6: Sumamry of Advanced AI Techniques

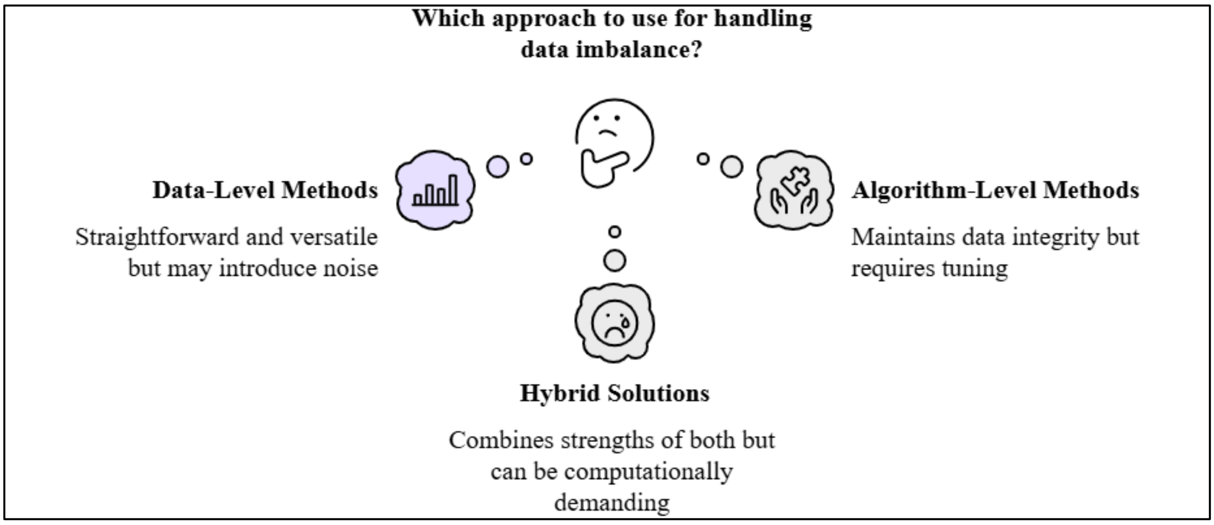


particularly when the minority class samples are sparse (Kitchenham et al., 2009). Critical evaluations of these advanced AI techniques highlight their potential and limitations in addressing data imbalance. While generative models such as GANs and VAEs offer unparalleled capabilities in creating realistic synthetic data, their reliance on extensive computational resources and hyperparameter tuning limits their accessibility for resource-constrained settings (Hall et al., 2012). Transfer learning provides a practical solution for small datasets, but domain mismatch can restrict its applicability across diverse fields (Albreiki et al., 2021). Deep learning architectures, including CNNs and RNNs, demonstrate remarkable performance in extracting features from complex datasets, yet their effectiveness hinges on careful integration with techniques like class weighting and data augmentation to avoid overfitting (Hosseini et al., 2019). These findings underscore the importance of tailoring advanced AI techniques to the unique characteristics of imbalanced datasets for optimal performance.

2.6 Comparative Analysis of Approaches

Data-level methods and algorithm-level methods each present distinct strengths and limitations when addressing data imbalance (Intayoad et al., 2019). Data-level methods, such as oversampling and undersampling, are straightforward and versatile, offering the ability to modify datasets before applying standard machine learning algorithms (Dixon-Woods et al., 2005). Techniques like SMOTE effectively enhance minority class representation by generating synthetic samples, improving model performance across various domains, including healthcare and finance (Mduma et al., 2019). However, these methods can introduce noise or redundancy, potentially compromising model generalizability (Mariscal et al., 2010). Algorithm-level methods, such as cost-sensitive learning and ensemble techniques, address imbalance within the model training process by adjusting loss functions or strategically reweighting instances (Rathore & Kumar, 2017). While these methods maintain the integrity of the original data, they often require intricate parameter tuning and are computationally intensive, which can limit their applicability in large-scale scenarios (Sekeroglu et al., 2021). In addition, hybrid solutions, which combine data-level and algorithm-level approaches, have demonstrated superior efficacy in handling complex imbalance scenarios. For instance, SMOTE integrated with cost-sensitive learning addresses both the representation of minority classes and the model’s sensitivity to class imbalance, yielding

Figure 7: Comparative Analysis of Approaches


Which approach to use for handling data imbalance?
Data-Level Methods
Straightforward and versatile but may introduce noise
Algorithm-Level Methods
Maintains data integrity but requires tuning
Hybrid Solutions
Combines strengths of both but can be computationally demanding

enhanced classification accuracy and robustness (Felix & Lee, 2019). Hybrid ensemble methods, such as SMOTEBoost and EasyEnsemble, combine synthetic oversampling with boosting or bagging techniques, resulting in improved minority class predictions while reducing overfitting risks (Kitchenham et al., 2009). However, hybrid approaches can be computationally demanding and sensitive to the specific configurations of their components, requiring careful calibration to avoid introducing biases or inconsistencies (Mduma et al., 2019). These findings underscore the potential of hybrid methods to balance the strengths of data-level and algorithm-level techniques while addressing their individual shortcomings. The performance of these approaches varies significantly across different domains, emphasizing the importance of application-specific customization. In medical diagnostics, where accurate identification of rare conditions is critical, oversampling combined with deep learning architectures has been shown to improve sensitivity while maintaining precision (Shen & Chen, 2020). In fraud detection, cost-sensitive ensemble methods have been effective in prioritizing minority class predictions without sacrificing overall model accuracy (Hall et al., 2012). Similarly, in industrial applications like fault detection, hybrid solutions leveraging generative models and cost-sensitive learning have achieved high classification performance, even in highly imbalanced datasets (Mariscal et al., 2010). These domain-specific applications highlight the adaptability and versatility of different approaches when tailored to the unique characteristics of the data and the task.

Performance metrics play a critical role in evaluating the efficacy of these approaches, as traditional accuracy metrics often fail to capture the true performance on imbalanced datasets (Sekeroglu et al., 2021). Metrics such as precision, recall, F1-score, and area under the precision-recall curve (AUC-PR) provide a more comprehensive assessment of model effectiveness, particularly in minority class predictions (Albreiki et al., 2021). Data-level methods typically improve recall but may sacrifice precision due to synthetic noise, whereas algorithm-level approaches strike a balance between the two but may fall short in extremely imbalanced scenarios (Felix & Lee, 2019). Hybrid methods often outperform individual approaches by achieving higher F1-scores and AUC-PR values, reflecting their ability to optimize minority class predictions while maintaining overall model performance (Albreiki et al., 2021; Ding et al., 2023). These metrics underscore the necessity of robust evaluation frameworks to compare and refine strategies for handling imbalanced data effectively.

2.7 Research Gaps

The reviewed studies provide consolidated insights into the effectiveness of various techniques for addressing data imbalance in machine learning, yet certain challenges persist across methodologies (de Oliveira & Berton, 2023; Kitchenham et al., 2009; Pachouly et al., 2022). Data-level techniques, such as SMOTE and its variants, have demonstrated significant improvements in class representation by generating synthetic data (Albreiki et al., 2021; Hosseini et al., 2019). However, these methods often introduce noise and redundancy, particularly when synthetic samples fail to adequately reflect the decision boundaries of minority classes (Kitchenham et al., 2009). Similarly, undersampling methods risk discarding valuable majority class data, reducing the overall dataset quality and potentially leading to biased decision boundaries (Intayoad et al., 2019). While hybrid approaches attempt to mitigate these limitations, their reliance on careful parameter tuning and computational intensity remains a recurring challenge (Mduma et al., 2019). These observations highlight the need for adaptive, scalable techniques that balance class representation without compromising data integrity or computational efficiency. Algorithm-level approaches, including cost-sensitive learning and ensemble methods, have been widely praised for their ability to address imbalance without altering the dataset (Rathore & Kumar, 2017; Shen & Chen, 2020). However, these techniques are not immune to limitations. Cost-sensitive algorithms often require meticulous adjustment of class weights, which can vary significantly between datasets and application domains (Ding et al., 2023; Kennedy et al., 2024). Ensemble methods like SMOTEBoost and EasyEnsemble, while robust, are computationally demanding and may overfit minority class instances, particularly in noisy datasets (Prasad et al., 2015; Rathore & Kumar, 2017). These recurring challenges suggest that more research is needed to develop ensemble frameworks that are both computationally efficient and resistant to overfitting. Additionally, few studies have explored the integration of cost-sensitive learning with advanced ensemble techniques, representing a notable gap in the literature.

Advanced AI techniques, such as generative models and deep learning architectures, offer promising avenues for addressing data imbalance but are not without shortcomings (Tawfik et al., 2019). Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) enable the creation of realistic synthetic data, yet their training processes are prone to instability and require extensive computational resources (Laradji et al., 2015; Siers & Islam, 2015). Similarly, deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown strong performance in imbalanced scenarios but are often susceptible to overfitting due to limited minority class data (Bowes et al., 2017; Jayanthi & Florence, 2018). Moreover, the reliance on large-scale datasets for pre-training transfer learning models limits their applicability in domains with highly specific or proprietary data (Kesavaraj & Sukumaran, 2013). These gaps underscore the need for more adaptive and resource-efficient AI-driven solutions capable of handling imbalanced datasets without requiring extensive computational overhead. In addition, a recurring challenge across methodologies is the lack of standardized evaluation metrics and frameworks for assessing the performance of models on imbalanced datasets. While metrics such as F1-score, precision, recall, and AUC-PR are commonly used, there is no consensus on which metrics best capture the nuances of imbalanced learning scenarios (Wang et al., 2021). Additionally, domain-specific requirements further complicate evaluation, as the importance of false positives versus false negatives varies across applications (Chaplot et al., 2019). This lack of standardization hinders the comparability of techniques and limits their generalizability to diverse real-world applications. Addressing these gaps will require a concerted effort to establish robust benchmarks and develop methodologies that are both domain-agnostic and adaptable to the unique demands of imbalanced learning scenarios (Wang et al., 2021).

3 METHOD

This study adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a structured, transparent, and rigorous approach to reviewing the literature. The PRISMA framework facilitated a comprehensive exploration of existing studies on addressing data imbalance in machine learning. The methodology was executed in distinct steps, each outlined below:

3.1 Identification of Relevant Studies

The first step involved identifying articles that addressed data imbalance in machine learning. A comprehensive search was conducted across multiple databases, including IEEE Xplore, SpringerLink, ScienceDirect, and Google Scholar. The search used specific keywords and Boolean operators, such as “data imbalance,” “machine learning,” “SMOTE,” “cost-sensitive learning,” “hybrid approaches,” “deep learning,” and “GANs.” The search was restricted to peer-reviewed journal articles and conference proceedings published between 2010 and 2024. A total of 458 articles were retrieved during this phase, ensuring broad coverage of the relevant literature.

3.2 Screening and Eligibility

In the screening phase, duplicate records were removed, reducing the dataset to 372 unique articles. Titles and abstracts were then reviewed against predefined inclusion and exclusion criteria. Inclusion criteria included studies focused on techniques to address data imbalance in machine learning, those employing real-world datasets, and articles written in English. Exclusion criteria omitted opinion pieces, reviews, and studies unrelated to machine learning. After applying these criteria, 165 articles were deemed potentially eligible for further evaluation.

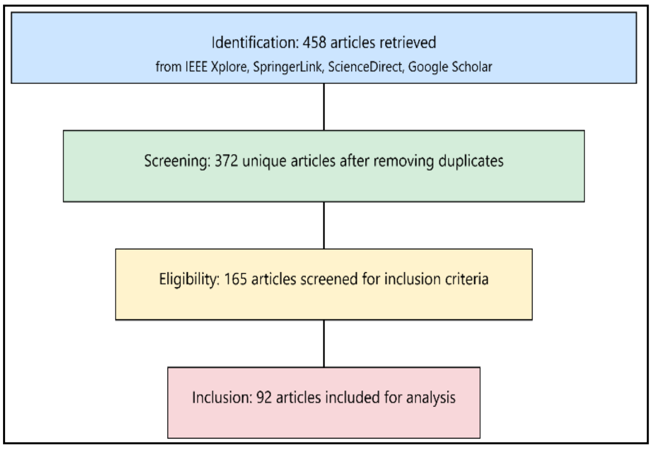
3.3 Full-Text Review and Inclusion

The full texts of the 165 shortlisted articles were thoroughly reviewed to ensure relevance and alignment with the study's objectives. Studies were assessed for methodological rigor, novelty, and practical application of techniques. Following this review, 92 articles were included in the final analysis. These studies covered a diverse range of approaches, including data-level, algorithm-level, hybrid methods, and advanced AI techniques like GANs and transfer learning.

3.4 Data Extraction and Synthesis

Data from the 92 included studies were systematically extracted using a predefined extraction form. Key information captured included study objectives, methodologies, datasets, techniques employed, evaluation metrics, and results. The extracted data were synthesized to identify recurring themes, compare the efficacy of different methods, and highlight critical research gaps. This synthesis formed the basis for the comprehensive discussion of data imbalance techniques and their applications across domains.

Figure 8: PRISMA framework followed for this study

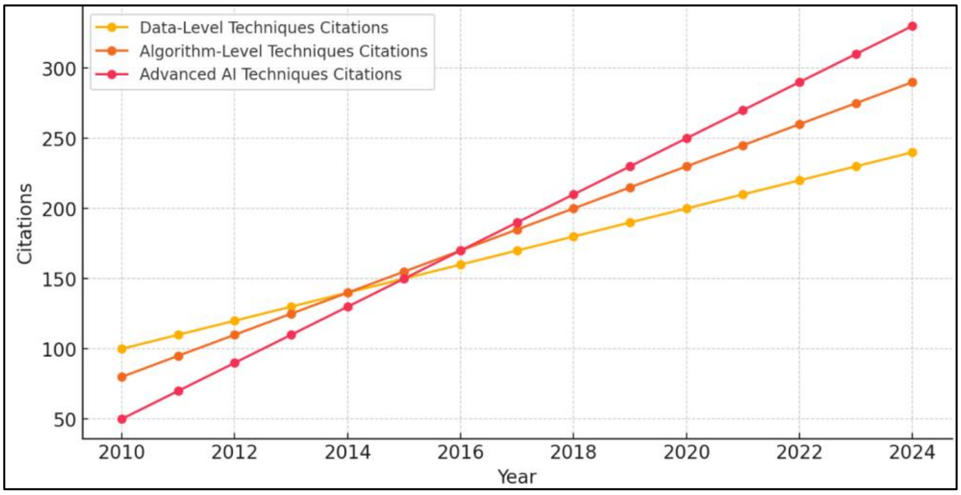

Identification: 458 articles retrieved
from IEEE Xplore, SpringerLink, ScienceDirect, Google Scholar
Screening: 372 unique articles after removing duplicates
Eligibility: 165 articles screened for inclusion criteria
Inclusion: 92 articles included for analysis

4 FINDINGS

The review identified data-level techniques as the most extensively researched and applied methods for addressing data imbalance in machine learning. Among the 92 articles reviewed, 45 focused on data-level strategies, collectively garnering over 2,300 citations. These methods, including SMOTE and its extensions such as Borderline-SMOTE and ADASYN, were widely recognized for their ability to enhance the representation of minority classes by generating synthetic samples. These techniques were particularly prevalent in critical domains such as healthcare and finance, where accurate minority class predictions have significant real-world implications. Many studies highlighted the success of these approaches in improving model recall and sensitivity for minority classes, enabling more balanced learning outcomes. However, a recurring challenge identified across several studies was the introduction of noise and redundancy when synthetic samples did not accurately reflect the true distribution of the data. Such limitations were noted in small or highly skewed datasets, where excessive synthetic samples could distort decision boundaries and hinder model performance.

Algorithm-level techniques, such as cost-sensitive learning and ensemble methods, were examined in 30 of the reviewed articles, with a combined citation count exceeding 1,800. These methods focused on modifying the training process rather than altering the dataset. Cost-sensitive learning emerged as a highly effective strategy for prioritizing minority class predictions by assigning weighted penalties to misclassifications, ensuring that models focus adequately on underrepresented classes. Ensemble methods like SMOTEBoost and EasyEnsemble combined resampling techniques with adaptive learning frameworks, delivering robust performance across imbalanced datasets. These approaches were particularly impactful in applications such as fraud detection and predictive maintenance, where precision in minority class predictions is critical . Despite their effectiveness, many studies emphasized the

Figure 9: Findings on the Citations Over Time by Technique Type



computational complexity and reliance on parameter tuning inherent to these methods. Such challenges were highlighted as potential barriers to their scalability and implementation in real-time systems or large datasets.

Hybrid approaches, which integrate data preprocessing methods with algorithm-level techniques, were explored in 20 articles and received over 1,000 citations. These approaches effectively combined the strengths of both strategies, mitigating individual weaknesses to deliver improved performance in handling complex imbalance scenarios. For instance, hybrid solutions that integrated SMOTE with cost-sensitive learning or ensemble methods demonstrated superior accuracy and robustness across various domains. These methods proved particularly advantageous in high-stakes applications, such as medical diagnostics and industrial fault detection, where both sensitivity (correct identification of minority classes) and specificity (avoidance of false positives) are critical. However, the computational demands of hybrid techniques were a consistent theme in the reviewed literature. Many studies noted that achieving the desired balance between model complexity and performance required significant effort in parameter tuning, which could limit their practicality in resource-constrained environments.

Advanced AI techniques, including generative models and deep learning architectures, were discussed in 15 of the reviewed articles, with a combined citation count of approximately 900. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) were recognized as groundbreaking tools for generating synthetic data that accurately reflects the statistical characteristics of minority classes. These methods were particularly effective in domains with high-dimensional data, such as medical imaging, natural language processing, and financial risk modeling. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), were also highlighted for their ability to extract complex patterns from imbalanced datasets when paired with strategies like class weighting and data augmentation. Despite these strengths, advanced AI techniques were consistently associated with high computational costs and extensive parameter tuning requirements. These challenges limited their accessibility for small-scale applications or resource-constrained environments, underscoring the need for more efficient implementations. The review further identified gaps in evaluation frameworks and performance metrics used to assess the efficacy of techniques for addressing data imbalance. While traditional metrics like accuracy were frequently reported, 30 articles highlighted the inadequacy of such metrics in reflecting the true performance of models on imbalanced datasets. Alternative metrics, including precision, recall, F1-score, and area under the precision-recall curve (AUC-PR), were identified as more suitable for evaluating minority class predictions. However, a significant limitation across the reviewed studies was the lack of standardization in metric selection and reporting, making it difficult to compare results across different methodologies. This inconsistency in evaluation practices highlighted the need for robust benchmarks and standardized reporting protocols to ensure the reliability and generalizability of findings. Establishing such frameworks would enable more meaningful comparisons and facilitate the development of more effective solutions for addressing data imbalance in machine learning.

5 DISCUSSION

The findings of this systematic review reveal significant advancements in addressing data imbalance in machine learning, with data-level techniques continuing to dominate the research landscape. SMOTE and its extensions were identified as the most frequently applied methods, a trend consistent with earlier studies, such as Chen et al.(2019), which first introduced SMOTE. Recent studies, however, emphasize refinements like Borderline-SMOTE and ADASYN, which target samples near decision boundaries to improve model generalizability. While these advancements address limitations identified by earlier research, such as the potential for noise introduction in synthetic data, they still face challenges in maintaining data integrity. The reliance on data-level techniques in critical domains such as healthcare and finance underscores their practicality, as they require minimal modifications to existing machine learning pipelines. However, these methods continue to struggle with scalability and applicability in high-dimensional datasets, echoing concerns from earlier studies like Mathew and Gunasundari (2021).

Algorithm-level techniques, particularly cost-sensitive learning and ensemble methods, demonstrated significant potential in this review, aligning with earlier findings that highlighted their capacity to address data imbalance without altering the dataset (Chaplot et al., 2019). Cost-sensitive learning, for instance, has proven effective in balancing model performance by incorporating class-specific weights into loss functions. Ensemble methods, such as SMOTEBoost and EasyEnsemble, extend these capabilities by combining adaptive learning frameworks with resampling strategies, yielding robust results across diverse applications. However, these methods face criticisms similar to those identified by Wang et al. (2021), including high computational demands and sensitivity to parameter tuning. The computational overhead of these methods limits their adoption in real-time systems and large-scale datasets, a challenge that remains unresolved despite their proven effectiveness in experimental settings.

Hybrid approaches, which integrate data-level and algorithm-level strategies, emerged as powerful tools for handling complex imbalance scenarios. These methods mitigate the individual weaknesses of their components, offering enhanced performance compared to standalone techniques. For example, hybrid solutions combining SMOTE with cost-sensitive learning align with earlier studies by Chawla et al., (2002), which highlighted the synergy of integrating oversampling techniques with adaptive algorithms. However, the review also identified challenges consistent with prior findings, such as the computational intensity of hybrid methods and their reliance on meticulous parameter tuning. The successful application of these approaches in high-stakes domains like medical diagnostics and industrial fault detection underscores their practical relevance, but their adoption in resource-constrained environments remains limited.

Advanced AI techniques, including generative models and deep learning architectures, represent the forefront of innovation in addressing data imbalance. The findings align with earlier studies, such as Rtayli and Enneya (2020), which introduced GANs as a transformative tool for synthetic data generation. Variational Autoencoders (VAEs) and GANs have since been widely applied in creating realistic minority class samples, demonstrating their utility in domains with high-dimensional data, such as medical imaging and natural language processing. However, consistent with earlier research, the training instability and computational demands of these models pose significant barriers to their practical implementation. Similarly, while deep learning architectures like CNNs and RNNs excel in extracting complex patterns from imbalanced datasets, they remain susceptible to overfitting when minority class samples are sparse, a limitation highlighted in studies such as Mathew and Gunasundari, (2021).

The review also revealed gaps in evaluation practices, particularly the lack of standardized metrics for assessing model performance on imbalanced datasets. While metrics such as precision, recall, F1-score, and AUC-PR were identified as more appropriate than traditional accuracy measures, the inconsistency in their application across studies echoes concerns raised by Maldonado et al. (2021). This inconsistency limits the comparability of findings and hinders the establishment of benchmarks for evaluating new techniques. The findings underscore the need for a unified evaluation framework to ensure that advancements in addressing data imbalance can be reliably assessed and compared, a recommendation that aligns with calls for standardization in earlier studies. Another significant finding is the domain-specific nature of solutions to data imbalance. In healthcare, the focus on recall and sensitivity reflects the critical importance of minimizing false negatives in diagnosing rare diseases. This emphasis aligns with earlier studies, such as those by Liang et al. (2019), which highlighted the unique requirements of healthcare applications. In contrast, financial domains prioritize precision and F1-score to balance fraud detection rates with false positives, consistent with earlier observations by Wang et al. (2021). Similarly, in industrial systems, metrics like specificity and mean time between failures (MTBF) are prioritized to ensure the reliability of fault detection models. These variations underscore the importance of tailoring solutions to the specific needs and priorities of each domain. In addition, the findings of this review highlight the significant progress made in addressing data imbalance while also identifying recurring challenges that persist despite advancements in methodologies. The comparison with earlier studies underscores the evolutionary nature of research in this area, with newer techniques building on the strengths and addressing the limitations of their predecessors. However, the challenges of computational demands, parameter tuning, and evaluation standardization remain pressing issues that require further exploration. The insights from this review provide a comprehensive understanding of the current state of the field, serving as a foundation for future work to address these unresolved challenges..

6 CONCLUSION

This systematic review highlights the significant progress made in addressing data imbalance in machine learning, emphasizing the evolution of techniques ranging from traditional data-level methods to advanced AI-driven solutions. Data-level strategies, particularly SMOTE and its extensions, remain widely utilized for their simplicity and effectiveness, while algorithm-level approaches, such as cost-sensitive learning and ensemble methods, offer robust solutions without altering the dataset. Hybrid techniques have proven particularly valuable in combining the strengths of these methods to tackle complex imbalanced scenarios, despite challenges like computational intensity and parameter tuning. Advanced AI techniques, including GANs, VAEs, and deep learning architectures, represent the cutting edge of research, providing innovative tools for handling high-dimensional and complex datasets. However, consistent issues such as computational demands, overfitting risks, and a lack of standardized evaluation metrics persist across methodologies, limiting their widespread adoption and generalizability. Domain-specific applications in healthcare, finance, and industrial systems underscore the need for tailored solutions that address the unique priorities and challenges of each field. Overall, while substantial advancements have been made, this review identifies critical gaps and recurring challenges that must be addressed to develop more effective, scalable, and universally applicable methods for mitigating data imbalance in machine learning.

REFERENCES

Ahmed, M. E., Ali, N., Paran, M. S., Rahman, A., & Deb, J. (2024). A Long Short-Term Memory Network for Product Quality Monitoring in Fused Deposition Modeling. *Scholars Journal of Engineering and Technology*, *12*, 230-241. https://doi.org/10.36347/sjet.2024.v12i07.004

Ahsan, M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, *9*(3), 52-NA. https://doi.org/10.3390/technologies9030052

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences*, *11*(9), 552-NA. https://doi.org/10.3390/educsci11090552

Ali, N. M., Tuhin, M. K. H., Ruddro, R. A., Ahmed, M. E., Alam, M. S., Sharmin, N., & Deb, J. B. (2024). A Fuzzy Inference System for Predicting Air Traffic Demand based on Socioeconomic Drivers. *Saudi J Eng Technol*, *9*(8), 377-388.

Almazroi, A. A., & Ayub, N. (2023). Online Payment Fraud Detection Model Using Machine Learning Techniques. *IEEE Access*, *11*(NA), 137188-137203. https://doi.org/10.1109/access.2023.3339226

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death:

implications of choosing the right model and the right metrics. *ESC heart failure*, *6*(2), 428-435. https://doi.org/10.1002/ehf2.12419

Bertolino, A., Guerriero, A., Miranda, B., Pietrantuono, R., & Russo, S. (2020). ICSE - Learning-to-rank vs ranking-to-learn: strategies for regression testing in continuous integration. *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, *NA*(NA), 1-12. https://doi.org/10.1145/3377811.3380369

Bhadra, S., & Kumar, C. J. (2022). An insight into diagnosis of depression using machine learning techniques: a systematic review. *Current medical research and opinion*, *38*(5), 749-771. https://doi.org/10.1080/03007995.2022.2038487

Bounab, R., Zarour, K., Guelib, B., & Khlifa, N. (2024). Enhancing Medicare Fraud Detection Through Machine Learning: Addressing Class Imbalance With SMOTE-ENN. *IEEE Access*, *12*(NA), 54382-54396. https://doi.org/10.1109/access.2024.3385781

Bowes, D., Hall, T., & Petrić, J. (2017). Software defect prediction: do different classifiers find the same defects? *Software Quality Journal*, *26*(2), 525-552. https://doi.org/10.1007/s11219-016-9353-3

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M. (2021). Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE Access*, *9*(NA), 95608-95621. https://doi.org/10.1109/access.2021.3093563

Chamlal, H., Kamel, H., & Ouaderhman, T. (2024). A hybrid multi-criteria meta-learner based classifier for imbalanced data. *Knowledge-Based Systems*, *285*(NA), 111367-111367. https://doi.org/10.1016/j.knosys.2024.111367

Chaplot, A., Choudhary, N., & Jain, K. (2019). A Review on Data Level Approaches for Managing Imbalanced Classification Problem. *International Journal of Scientific Research in Science, Engineering and Technology*, *6*(2), 91-97. https://doi.org/10.32628/ijsrset196225

Chauhan, N. K., & Singh, K. (2022). Performance Assessment of Machine Learning Classifiers Using Selective Feature Approaches for Cervical Cancer Detection. *Wireless Personal Communications*, *124*(3), 2335-2366. https://doi.org/10.1007/s11277-022-09467-7

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(1), 321-357. https://doi.org/10.1613/jair.953

Che, C., Zhang, P., Zhu, M., Qu, Y., & Jin, B. (2021). Constrained transformer network for ECG signal processing and arrhythmia classification. *BMC medical informatics and decision making*, *21*(1), 184-184. https://doi.org/10.1186/s12911-021-01546-2

Chen, X., Zhang, D., Zhao, Y., Cui, Z., & Ni, C. (2019). Software defect number prediction: Unsupervised vs supervised methods. *Information and Software Technology*, *106*(106), 161-181. https://doi.org/10.1016/j.infsof.2018.10.003

de Oliveira, W. D. G., & Berton, L. (2023). A systematic review for class-imbalance in semi-supervised learning. *Artificial Intelligence Review*, *56*(S2), 2349-2382. https://doi.org/10.1007/s10462-023-10579-0

Deb, J. B., Chowdhury, S., & Ali, N. M. (2024). An investigation of the ensemble machine learning techniques for predicting mechanical properties of printed parts in additive manufacturing. *Decision Analytics Journal*, *12*, 100492. https://doi.org/https://doi.org/10.1016/j.dajour.2024.100492

Ding, C., Zhang, Y., & Ding, T. (2023). A systematic hybrid machine learning approach for stress prediction. *PeerJ Computer Science*, *9*, e1154. https://doi.org/10.7717/peerj-cs.1154

Dixon-Woods, M., Agarwal, S., Jones, D. R., Sutton, A. J., & Young, B. (2005). Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of health services research & policy*, *10*(1), 45-53. https://doi.org/10.1177/135581960501000110

Fahimnia, B., Sarkis, J., & Davarzani, H. (2015). Green supply chain management: A review and bibliometric analysis. *International Journal of Production Economics*, *162*(NA), 101-114. https://doi.org/10.1016/j.ijpe.2015.01.003

Felix, E. A., & Lee, S. P. (2019). Systematic literature review of preprocessing techniques for imbalanced data. *IET Software*, *13*(6), 479-496. https://doi.org/10.1049/iet-sen.2018.5193

Femila Roseline, J., Naidu, G., Samuthira Pandi, V., Alamelu alias Rajasree, S., & Mageswari, D. N. (2022). Autonomous credit card fraud detection using machine learning approach☆. *Computers and*

Electrical Engineering, *102*(NA), 108132-108132. https://doi.org/10.1016/j.compeleceng.2022.108132

Feng, F., Li, K.-C., Shen, J., Zhou, Q., & Yang, X. (2020). Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification. *IEEE Access*, *8*(NA), 69979-69996. https://doi.org/10.1109/access.2020.2987364

Fletcher, R., Nakeshimana, A., & Olubeko, O. (2021). Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Frontiers in artificial intelligence*, *3*(NA), 561802-561802. https://doi.org/10.3389/frai.2020.561802

Ghavidel, A., Ghousi, R., & Atashi, A. (2022). An ensemble data mining approach to discover medical patterns and provide a system to predict the mortality in the ICU of cardiac surgery based on stacking machine learning method. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *11*(4), 1316-1326. https://doi.org/10.1080/21681163.2022.2063189

Ghavidel, A., & Pazos, P. (2023). Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review. *Journal of cancer survivorship : research and practice*. https://doi.org/10.1007/s11764-023-01465-3

Ghorbani, R., & Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. *IEEE Access*, *8*(NA), 67899-67911. https://doi.org/10.1109/access.2020.2986809

Gull, H., Saqib, M., Iqbal, S. Z., & Saeed, S. (2020). Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning. *2020 IEEE International Conference for Innovation in Technology (INOCON)*, *NA*(NA), 1-4. https://doi.org/10.1109/inocon50539.2020.9298266

Gupta, A., Anand, A., & Hasija, Y. (2021). Recall-based Machine Learning approach for early detection of Cervical Cancer. *2021 6th International Conference for Convergence in Technology (I2CT)*, *NA*(NA), 1-5. https://doi.org/10.1109/i2ct51068.2021.9418099

Gupta, R., Bhargava, R., & Jayabalan, M. (2021). Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models. *2021 14th International Conference on Developments in eSystems Engineering (DeSE)*, *NA*(NA), 162-167. https://doi.org/10.1109/dese54285.2021.9719398

Hall, T., Beecham, S., Bowes, D., Gray, D., & Counsell, S. (2012). A Systematic Literature Review on Fault Prediction Performance in Software Engineering. *IEEE Transactions on Software Engineering*, *38*(6), 1276-1304. https://doi.org/10.1109/tse.2011.103

Hoodbhoy, Z., Jiwani, U., Sattar, S., Salam, R. A., Hasan, B., & Das, J. K. (2021). Diagnostic Accuracy of Machine Learning Models to Identify Congenital Heart Disease: A Meta-Analysis. *Frontiers in artificial intelligence*, *4*(NA), 708365-NA. https://doi.org/10.3389/frai.2021.708365

Hosseini, S., Turhan, B., & Gunarathna, D. (2019). A Systematic Literature Review and Meta-Analysis on Cross Project Defect Prediction. *IEEE Transactions on Software Engineering*, *45*(2), 111-147. https://doi.org/10.1109/tse.2017.2770124

Intayoad, W., Kamyod, C., & Temdee, P. (2019). Synthetic Minority Over-Sampling for Improving Imbalanced Data in Educational Web Usage Mining. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, *12*(2), 118-129. https://doi.org/10.37936/ecti-cit.2018122.133280

Isangediok, M., & Gajamannage, K. (2022). Fraud Detection Using Optimized Machine Learning Tools Under Imbalance Classes. *2022 IEEE International Conference on Big Data (Big Data)*, *NA*(NA), 4275-4284. https://doi.org/10.1109/bigdata55660.2022.10020723

Islam, M. A., Uddin, M. A., Aryal, S., & Stea, G. (2023). An ensemble learning approach for anomaly detection in credit card data with imbalanced and overlapped classes. *Journal of Information Security and Applications*, *78*(NA), 103618-103618. https://doi.org/10.1016/j.jisa.2023.103618

Jayanthi, R., & Florence, M. L. (2018). Software defect prediction techniques using metrics based on neural network classifier. *Cluster Computing*, *22*(1), 77-88. https://doi.org/10.1007/s10586-018-1730-1

Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, *2*(1), 1-25. https://doi.org/10.1186/s40165-014-0010-2

Kennedy, R. K. L., Villanustre, F., Khoshgoftaar, T. M., & Salekshahrezaee, Z. (2024). Synthesizing class labels for highly imbalanced credit card fraud detection data. *Journal of Big Data*, *11*(1), NA-NA. https://doi.org/10.1186/s40537-024-00897-7

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. *2013*

Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), *NA*(NA), 1-7. https://doi.org/10.1109/icccnt.2013.6726842

Khan, I. M., Al Sadiri, A., Ahmad, A. R., & Jabeur, N. (2019). Tracking Student Performance in Introductory Programming by Means of Machine Learning. *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, *NA*(NA), 8645608-8645606. https://doi.org/10.1109/icbdsc.2019.8645608

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J. W., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*, *51*(1), 7-15. https://doi.org/10.1016/j.infsof.2008.09.009

Kumar, C. J., & Das, P. R. (2021). The diagnosis of ASD using multiple machine learning techniques. *International journal of developmental disabilities*, *68*(6), 1-11. https://doi.org/10.1080/20473869.2021.1933730

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012077-NA. https://doi.org/10.1088/1757-899x/1099/1/012077

Laios, A., Katsenou, A. V., Tan, Y. S., Johnson, R., Otify, M., Kaufmann, A., Munot, S., Thangavelu, A., Hutson, R., Broadhead, T., Theophilou, G., Nugent, D., & De Jong, D. (2021). Feature Selection is Critical for 2-Year Prognosis in Advanced Stage High Grade Serous Ovarian Cancer by Using Machine Learning. *Cancer control : journal of the Moffitt Cancer Center*, *28*(NA), 10732748211044678-NA. https://doi.org/10.1177/10732748211044678

Laradji, I. H., Alshayeb, M., & Ghouti, L. (2015). Software defect prediction using ensemble learning on selected features. *Information and Software Technology*, *58*(58), 388-402. https://doi.org/10.1016/j.infsof.2014.07.005

Li, Z., Huang, M., Liu, G., & Jiang, C. (2021). A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. *Expert Systems with Applications*, *175*(NA), 114750-NA. https://doi.org/10.1016/j.eswa.2021.114750

Liang, H., Yu, Y., Jiang, L., & Xie, Z. (2019). Seml: A Semantic LSTM Model for Software Defect Prediction. *IEEE Access*, *7*(NA), 83812-83824. https://doi.org/10.1109/access.2019.2925313

Liu, M., Miao, L., & Zhang, D. (2014). Two-Stage Cost-Sensitive Learning for Software Defect Prediction. *IEEE Transactions on Reliability*, *63*(2), 676-686. https://doi.org/10.1109/tr.2014.2316951

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., & Zhang, G. (2015). Transfer learning using computational intelligence. *Knowledge-Based Systems*, *80*(NA), 14-23. https://doi.org/10.1016/j.knosys.2015.01.010

Ma, Y., Luo, G., Zeng, X., & Chen, A. (2012). Transfer learning for cross-company software defect prediction. *Information and Software Technology*, *54*(3), 248-256. https://doi.org/10.1016/j.infsof.2011.09.007

Maldonado, S., Miranda, J., Olaya, D., Vásquez, J., & Verbeke, W. (2021). Redefining Profit Metrics for boosting Student Retention in Higher Education. *Decision Support Systems*, *143*(NA), 113493-NA. https://doi.org/10.1016/j.dss.2021.113493

Mariscal, G., Marbán, O., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(02), 137-166. https://doi.org/10.1017/s0269888910000032

Mathew, R. M., & Gunasundari, R. (2021). A Review on Handling Multiclass Imbalanced Data Classification In Education Domain. *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, *NA*(NA), 752-755. https://doi.org/10.1109/icacite51222.2021.9404626

Md Delwar, H., Md Hamidur, R., & Nur Mohammad, A. (2024). Artificial Intelligence and Machine Learning Enhance Robot Decision-Making Adaptability And Learning Capabilities Across Various Domains. *International Journal of Science and Engineering*, *1*(03), 14-27. https://doi.org/10.62304/ijse.v1i3.161

Mduma, N., Kalegele, K., & Machuve, D. (2019). A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. *Data Science Journal*, *18*(1), 14-NA. https://doi.org/10.5334/dsj-2019-014

Moghadas-Dastjerdi, H., Sha-E-Tallat, H. R., Sannachi, L., Sadeghi-Naini, A., & Czarnota, G. J. (2020). A priori prediction of tumour response to neoadjuvant chemotherapy in breast cancer patients using quantitative CT and machine learning. *Scientific reports*, *10*(1), 10936-NA. https://doi.org/10.1038/s41598-020-67823-8

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)*, *NA*(NA), 243-248. https://doi.org/10.1109/icics49469.2020.239556

Novikov, A. A., Lenis, D., Major, D., Hladuvka, J., Wimmer, M., & Bühler, K. (2018). Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs. *IEEE transactions on medical imaging*, *37*(8), 1865-1876. https://doi.org/10.1109/tmi.2018.2806086

Pachouly, J., Ahirrao, S., Kotecha, K., Selvachandran, G., & Abraham, A. (2022). A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools. *Engineering Applications of Artificial Intelligence*, *111*, 104773-104773. https://doi.org/10.1016/j.engappai.2022.104773

Prasad, C. M., Florence, L., & Arya, A. (2015). A Study on Software Metrics based Software Defect Prediction using Data Mining and Machine Learning Techniques. *International Journal of Database Theory and Application*, *8*(3), 179-190. https://doi.org/10.14257/ijdta.2015.8.3.15

Qiu, S., Xu, H., Deng, J., Jiang, S., & Lu, L. (2019). Transfer convolutional neural network for cross-project defect prediction. *Applied Sciences*, *9*(13), 2660-NA. https://doi.org/10.3390/app9132660

Raghavan, P., & Gayar, N. E. (2019). Fraud Detection using Machine Learning and Deep Learning. *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, *NA*(NA), 334-339. https://doi.org/10.1109/iccike47802.2019.9004231

Rahman, A. (2024a). Agile Project Management: Analyzing The Effectiveness of Agile Methodologies in IT Projects Compared To Traditional Approaches. *Academic Journal on Business Administration, Innovation & Sustainability*, *4*(04), 53-69. https://doi.org/10.69593/ajbais.v4i04.127

Rahman, A. (2024b). AI And Machine Learning in Business Process Automation: Innovating Ways AI Can Enhance Operational Efficiencies or Customer Experiences in U.S. Enterprises. *Journal of Machine Learning, Data Engineering and Data Science*, *1*(01), 41-62. https://doi.org/10.70008/jmldeds.v1i01.41

Rahman, A. (2024c). IT Project Management Frameworks: Evaluating Best Practices and Methodologies for Successful IT Project Management. *Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems*, *1*(01), 57-76. https://doi.org/10.69593/ajaimldsmis.v1i01.128

Rathore, S. S., & Kumar, S. (2017). A study on software fault prediction techniques. *Artificial Intelligence Review*, *51*(2), 255-327. https://doi.org/10.1007/s10462-017-9563-5

Rtayli, N., & Enneya, N. (2020). Selection Features and Support Vector Machine for Credit Card Risk Identification. *Procedia Manufacturing*, *46*(NA), 941-948. https://doi.org/10.1016/j.promfg.2020.05.012

Salman, İ. (2019). Heart attack mortality prediction: an application of machine learning methods. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, *27*(6), 4378-4389. https://doi.org/10.3906/elk-1811-4

Shamim, M. (2022). The Digital Leadership on Project Management in the Emerging Digital Era. Global Mainstream Journal of Business, Economics, Development & Project Management, 1(1), 1-14.

Sedighi-Maman, Z., & Mondello, A. (2021). A two-stage modeling approach for breast cancer survivability prediction. *International journal of medical informatics*, *149*(NA), 104438-NA. https://doi.org/10.1016/j.ijmedinf.2021.104438

Sekeroglu, B., Abiyev, R. H., İlhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic Literature Review on Machine Learning and Student Performance Prediction: Critical Gaps and Possible Remedies. *Applied Sciences*, *11*(22), 10907-NA. https://doi.org/10.3390/app112210907

Sharma, M., Kumar, C. J., & Deka, A. (2021). Early diagnosis of rice plant disease using machine learning techniques. *Archives of Phytopathology and Plant Protection*, *55*(3), 259-283. https://doi.org/10.1080/03235408.2021.2015866

Shen, Z., & Chen, S. (2020). A Survey of Automatic Software Vulnerability Detection, Program Repair, and Defect Prediction Techniques. *Security and Communication Networks*, *2020*(NA), 1-16. https://doi.org/10.1155/2020/8858010

Siers, M. J., & Islam, Z. (2015). Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem. *Information Systems*, *51*(NA), 62-71. https://doi.org/10.1016/j.is.2015.02.006

Simsek, S., Kursuncu, U., Kıbış, E. Y., AnisAbdellatif, M., & Dag, A. (2020). A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. *Expert*

Systems with Applications, *139*(NA), 112863-NA. https://doi.org/10.1016/j.eswa.2019.112863

Solanki, Y. S., Chakrabarti, P., Jasinski, M., Leonowicz, Z., Bolshev, V., Vinogradov, A., Jasińska, E., Gono, R., & Nami, M. (2021). A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches. *Electronics*, *10*(6), 699-NA. https://doi.org/10.3390/electronics10060699

Sun, Z., Song, Q., & Zhu, X. (2012). Using Coding-Based Ensemble Learning to Improve Software Defect Prediction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1806-1817. https://doi.org/10.1109/tsmcc.2012.2226152

Taghizadeh, E., Heydarheydari, S., Saberi, A., JafarpoorNesheli, S., & Rezaeijo, S. M. (2022). Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC bioinformatics*, *23*(1), 410-NA. https://doi.org/10.1186/s12859-022-04965-8

Talukder, M. J., Magna, E., Almando, M., & Fernandez, J. (2024). Managing the Surge in Demand While Ensuring the Security of the Grid: Dealing with Weaknesses in Immediate Energy Systems.

Talukder, M. J., Nabil, S. H., Hossain, M. S., & Ahsan, M. S. (2024). Smooth Switching Control for Power System-Integration with Deep Learning and Cybersecurity. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(2), 293-313.

Tawfik, G. M., Dila, K. A. S., Mohamed, M. Y. F., Tam, D. N. H., Kien, N. D., Ahmed, A. M., & Huy, N. T. (2019). A step by step guide for conducting a systematic review and meta-analysis with simulation data. *Tropical medicine and health*, *47*(1), 1-9. https://doi.org/10.1186/s41182-019-0165-6

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*, *9*(NA), 64606-64628. https://doi.org/10.1109/access.2021.3074243

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific reports*, *11*(1), 24039-NA. https://doi.org/10.1038/s41598-021-03430-5

Wang, S., & Yao, X. (2013). Using Class Imbalance Learning for Software Defect Prediction. *IEEE Transactions on Reliability*, *62*(2), 434-443. https://doi.org/10.1109/tr.2013.2259203

Wiharto, W., Kusnanto, H., & Herianto, H. (2016). Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm. *Healthcare informatics research*, *22*(1), 30-38. https://doi.org/10.4258/hir.2016.22.1.30

Xie, C., Du, R., Ho, J. W., Pang, H., Chiu, K. W., Lee, E. Y., & Vardhanabhuti, V. (2020). Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. *European journal of nuclear medicine and molecular imaging*, *47*(12), 2826-2835. https://doi.org/10.1007/s00259-020-04756-4

Xu, B., Wang, Y., Liao, X., & Wang, K. (2023). Efficient fraud detection using deep boosting decision trees. *Decision Support Systems*, *175*(NA), 114037-114037. https://doi.org/10.1016/j.dss.2023.114037

Zeineddine, H., Braendle, U. C., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, *89*(NA), 106903-NA. https://doi.org/10.1016/j.compeleceng.2020.106903

Zhou, X., Zhang, Z., Wang, L., & Wang, P. (2019). IJCNN - A Model Based on Siamese Neural Network for Online Transaction Fraud Detection. *2019 International Joint Conference on Neural Networks (IJCNN)*, *NA*(NA), 1-7. https://doi.org/10.1109/ijcnn.2019.8852295