

From Numbers to Narratives: Efficient Language Model-Based Detection for Safety-Critical Minority Classes

Anonymous ACL submission

Abstract

Safety-critical classification tasks face a persistent challenge: traditional models achieve high overall accuracy but inadequate performance on critical minority classes. We introduce a “numbers to narratives” framework that transforms tabular data into contextually rich descriptions, enabling language models to leverage pre-trained knowledge for minority class detection. Our approach integrates structured verbalization, linguistically-informed augmentation, and parameter-efficient fine-tuning to address the “minority class blind spot” in high-consequence domains. Using a significantly more efficient model architecture than existing approaches, our framework achieves superior minority class F1-scores: 78.76% for machine failures (+7.42 points over XGBoost), 65.87% for at-risk students (+12.12 points over MLP), and 32.00% for semiconductor failures (+1.01 points over XGBoost, despite 14:1 class imbalance). Our approach also improves overall accuracy by up to 22.43% in five of six datasets while maintaining computational feasibility. Ablation studies confirm that narrative-based verbalization enables effective reasoning about tabular data by contextualizing abstract numerical features. This work provides a practical, resource-efficient approach for enhancing minority class performance in safety-critical domains.

1 Introduction

Safety-critical classification tasks present persistent challenges across diverse domains such as healthcare, manufacturing, and transportation. Tabular datasets from the UCI Machine Learning Repository (Dua and Graff, 2017) frequently exhibit significant class imbalance, where critical events of interest (e.g., machine failures, medical complications) represent a small fraction of instances. Traditional machine learning (ML) algorithms and neural architectures applied to these datasets often

achieve misleadingly high overall accuracy while substantially underperforming on minority classes (Fernández et al., 2018; Provost and Fawcett, 2013). This performance disparity is illustrated in our analysis of the UCI AI4I predictive maintenance dataset, where XGBoost achieves 97.75% overall accuracy but only 71.34% F1-score for the critical machine failure class—a gap that could translate to missed detection of impending equipment failures with significant operational consequences (Johnson and Khoshgoftaar, 2019).

This “minority class blind spot” creates a troubling disconnect between reported model performance and practical utility in high-consequence decision domains. Even with optimal hyperparameter tuning, traditional ML algorithms such as Random Forest (Breiman, 2001), Support Vector Machines (SVMs) (Hearst et al., 1998), and XGBoost (Chen and Guestrin, 2016) struggle with imbalanced class distributions due to their optimization for aggregate metrics (He and Garcia, 2009). Similarly, deep learning (DL) architectures, including Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs), frequently underperform on small or imbalanced datasets, particularly when feature-to-sample ratios are unfavorable (Buda et al., 2018).

Existing remediation strategies for class imbalance fall into two categories, each with significant limitations. *Data-level approaches* such as Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) generate synthetic samples but introduce statistical noise and distort feature distributions, particularly in datasets with complex feature interactions. *Algorithm-level approaches* such as cost-sensitive learning (Elkan, 2001) improve minority class detection but require domain expertise to set appropriate cost matrices. These limitations are especially pronounced in tabular data, where numerical features carry implicit semantic meaning that conventional tech-

niques struggle to preserve (Kotsiantis et al., 2006). Recent approaches leveraging Transformer-based (Vaswani et al., 2017) language models (LMs) show promise for tabular data tasks. TAPAS (Herzig et al., 2020) incorporates table-specific embeddings and positional encodings to capture structural relationships for question-answering tasks, while TabLLM (Hegselmann et al., 2023) serializes tabular data into natural language strings for classification. However, these models are not designed for minority class performance in safety-critical domains and may not fully capture domain-specific contextual knowledge without customization.

To overcome these shortcomings, we propose a novel “numbers to narratives” framework that transforms tabular data into contextually rich natural language descriptions. This approach enables language models to leverage their pre-trained general knowledge about the world for safety-critical classification tasks where minority class detection is paramount. Our framework integrates three complementary components: (1) structured verbalization, which converts numerical instances into semantically coherent text (e.g., *Patient aged 42 with a history of hypertension and elevated glucose levels showing early signs of retinopathy* instead of *Age: 42, Hypertension: Yes, Glucose: 182, Retinopathy: Early*), preserving feature relationships and domain context; (2) linguistically-informed minority class augmentation, which generates context-aware synthetic samples while maintaining causal dependencies to mitigate class imbalance without distorting feature distributions; and (3) parameter-efficient fine-tuning using quantized low-rank adaptation (QLoRA) (Dettmers et al., 2023) to adapt pre-trained language models to domain-specific tabular tasks with minimal computational overhead.

Unlike previous approaches that rely on large LMs (or LLMs) such as TabLLM (Hegselmann et al., 2023) (11B parameters) and TAPAS (Herzig et al., 2020) (BERT-large, 340M parameters), our framework achieves superior performance with significantly lower computational requirements. By utilizing a 66M-parameter DistilBERT model (Sanh et al., 2020) with parameter-efficient fine-tuning, we reduce resource needs by 5-160× while improving minority class detection—making our approach both more effective and more accessible for real-world deployment in resource-constrained environments. This framework addresses **three critical research questions (RQs)**:

1. **RQ1:** How do LMs with verbalized tabular inputs compare to conventional ML and DL models across datasets with varying class balance, scale, and feature complexity?
2. **RQ2:** Can linguistically-informed augmentation in LM-based approaches outperform data-level methods such as SMOTE in improving minority class performance for safety-critical domains?
3. **RQ3:** How do different verbalization strategies and instruction-based fine-tuning impact LM performance in structured data classification tasks?

Our comprehensive evaluation across six UCI datasets (Dua and Graff, 2017) varying in class balance, scale, and domain demonstrates that the “numbers to narratives” framework significantly improves both minority class detection and overall accuracy. The approach achieves notable gains in safety-critical scenarios: 78.76% F1-score for detecting machine failures in AI4I (7.42 percentage points over XGBoost), 65.87% for identifying at-risk students (12.12 points over MLP), and 32.00% for semiconductor failures in SECOM (1.01 points over XGBoost despite extreme 14:1 class imbalance). In five of six datasets, our approach enhances overall accuracy by 1.50–22.43%. For the SECOM dataset, we make a deliberate trade-off: traditional models achieve ~93% accuracy but effectively miss critical failure cases ($F1 \leq 0.09\%$), while our approach reaches 67% accuracy but attains meaningful minority class detection (32% F1)—directly addressing the “minority class blind spot” that renders seemingly high-performing models ineffective for safety-critical applications.

This work contributes to the growing body of NLP research on cross-modal applications, where natural language understanding capabilities enhance performance on structured data (Bommasani et al., 2022). By transforming tabular data into contextually rich descriptions, our approach enables language models to reason effectively about tabular instances while maintaining computational efficiency.

Our main contributions include:

- I. A novel “numbers to narratives” framework that transforms tabular data into contextually rich natural language descriptions, enabling LMs to leverage pre-trained knowledge for safety-critical classification tasks.
- II. A linguistically-informed minority class aug-

mentation approach that preserves semantic relationships while addressing class imbalance, outperforming data-level methods such as SMOTE by generating context-aware synthetic samples.

- III. Significant performance improvements in minority class detection (up to +12.12 points F1-score) and overall accuracy (up to +22.43%) across diverse datasets, even with extreme class imbalance.
- IV. A computationally efficient approach requiring 5-160× fewer resources than existing LM-based tabular methods (66M parameters vs. 340M-11B), enabling training on a single GPU in under an hour.
- V. Actionable insights from ablation studies quantifying the impact of verbalization strategies, augmentation techniques, and few-shot learning for applying LMs to safety-critical tabular data classification.

2 Related Work

Classical ML and DL for Tabular Data Both conventional ML models (k-NN, Decision Trees, Random Forests, SVMs, XGBoost) (Dua and Graff, 2017; Chen and Guestrin, 2016) and specialized deep learning architectures (TabNet (Arik and Pfister, 2020), NODE (Popov et al., 2019)) face persistent challenges with tabular data in safety-critical domains. Studies on unbalanced datasets like AI4I often exhibit a substantial gap between aggregate evaluation metrics and minority class F1-scores (Johnson and Khoshgoftaar, 2019), while balanced multi-class tasks like Glass Identification show precision/recall below 65% for specific classes (McCann and Johnston, 2008). Despite architectural advances, neural approaches frequently underperform on small datasets due to overfitting (Goodfellow et al., 2016; Kotsiantis et al., 2006) and prioritize aggregate metrics over minority class performance. Recent benchmarks confirm that well-tuned tree ensembles still outperform specialized neural architectures on many tabular tasks, particularly those with complex feature interactions and limited samples, highlighting the persistent “minority class blind spot” that undermines practical utility in safety-critical applications (Grinsztajn et al., 2022).

Addressing Class Imbalance Imbalance strategies include data-level (e.g., SMOTE (Chawla et al., 2002), <70% recall on Gas (Vergara et al., 2012);

ADASYN (He et al., 2008)), algorithm-level (cost-sensitive learning (Elkan, 2001), ensembles like SMOTEBoost (Chawla et al., 2003), RUSBoost (Seiffert et al., 2010)), and hybrid methods. These often introduce noise or lose information (Provost and Fawcett, 2013), failing to capture semantic relationships. These approaches struggle to generalize across datasets with varying characteristics, highlighting the need for context-aware augmentation strategies that preserve semantic integrity.

Language Models for Structured Data Pre-trained LLMs such as TAPAS (Herzig et al., 2020) and TaBERT (Yin et al., 2020) target table QA (>75% accuracy), while TabLLM (Hegselmann et al., 2023), TableFormer (Yang et al., 2022), and TUTA (Wang et al., 2021) focus on classification but neglect imbalance. Li et al. (Li et al., 2024) and Borisov et al. (Borisov et al., 2024) enhance semantic learning, yet overlook minority detection. Our framework advances with context-aware verbalization, semantic augmentation, and fine-tuning for safety-critical tasks.

Few-Shot Learning and Efficient Fine-Tuning Few-shot learning excels with in-context examples (Brown et al., 2020), but is sensitive to selection (Min et al., 2022) and limited for tabular imbalance (Wei et al., 2022). QLoRA (Dettmers et al., 2023) and LoRA (Hu et al., 2021) offer efficient fine-tuning (<1% parameters), yet tabular safety applications are underexplored (Hegselmann et al., 2023).

To our knowledge, no prior work integrates structured verbalization, linguistically-informed augmentation, and efficient fine-tuning for safety-critical minority class detection in tabular data—a gap our approach bridges while addressing practical dataset constraints.

3 Methodology

This section presents our “numbers to narratives” framework for safety-critical classification through language model verbalization of tabular data, designed to address the limitations of traditional ML and DL approaches on minority classes. Our methodology emphasizes robustness across diverse dataset characteristics: class balance (balanced vs. unbalanced), scale (small vs. large), feature complexity, and task type (binary vs. multi-class). We evaluate the framework on six UCI datasets (Dua and Graff, 2017) with varying characteristics, as summarized in Table 1.

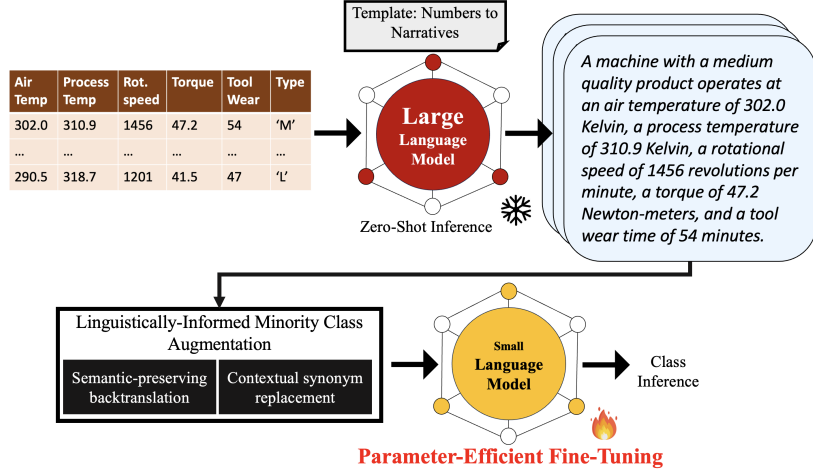


Figure 1: The “numbers to narratives” framework: transforming tabular data into natural language, augmenting minority classes, and efficiently fine-tuning a compact language model.

Table 1: Dataset Characteristics

Dataset	Size	Features	Classes	Ratio
AI4I 2020	10,000	6	2	~24:1
Glass	214	9	6	varies
Student	395	33	2	~3:1 to 5:1
Gas	13,910	128	6	balanced
Mammographic	961	5	2	~1:1
SECOM	1,567	590	2	~14:1

Our framework (illustrated in Figure 1) consists of three sequential components that transform tabular data into effective safety-critical classifiers while maintaining computational efficiency. Each component builds upon the previous one to address specific challenges in minority class detection.

3.1 Structured Verbalization of Tabular Data

First, we transform numerical tabular instances into natural language descriptions using a context-aware verbalization approach. This process leverages a large generative language model—ChatGPT-4o (OpenAI, 2024)—in a zero-shot manner to convert abstract feature vectors into semantically rich textual representations. The verbalization follows a structured template (detailed in Appendix A.4.1) that provides three levels of context:

1. **Domain context:** Dataset overview, task description, and feature explanations
2. **Feature semantics:** Natural language descriptions of feature meanings and relationships
3. **Instance-specific narration:** Coherent narrative integrating all feature values

For each dataset, we create a standardized prompt template to ensure consistent verbalization patterns. This template maps numerical and categorical features to contextually appropriate linguistic expressions, preserving the semantic rela-

tionships between features. For example, an AI4I instance with numeric values [302.0, 310.9, 1456, 47.2, 54, ‘M’] is transformed into narrative text: “A machine with a medium quality product operates at an air temperature of 302.0 Kelvin, a process temperature of 310.9 Kelvin, a rotational speed of 1456 revolutions per minute, a torque of 47.2 Newton-meters, and a tool wear time of 54 minutes.” This rich description contextualizes the abstract numeric features within their physical meaning and relationships.

This verbalization approach transforms abstract feature spaces into human-interpretable narratives, enabling language models to apply their pre-trained knowledge about real-world relationships to the classification task (Brown et al., 2020; Wei et al., 2022). Unlike previous approaches such as TabLLM (Hegselmann et al., 2023) that use simple “feature: value” mappings, our method generates cohesive narratives that preserve causal and semantic relationships between features.

3.2 Linguistically-Informed Minority Class Augmentation

Second, we address class imbalance while preserving semantic integrity through linguistically-informed augmentation of the verbalized data. For binary tasks (e.g., AI4I, SECOM), we match minority to majority class size; for multi-class tasks (e.g., Glass), we balance underrepresented classes. The augmentation pipeline includes:

1. **Semantic-preserving backtranslation:** We translate verbalized instances from English to German and back to English using the facebook/wmt19-en-de and facebook/wmt19-de-en models (both 270M parameters) (Ng et al.,

2019) in a zero-shot manner. This process generates linguistic variation while retaining core meaning, building on established backtranslation methods (Sennrich et al., 2016).

2. **Contextual synonym replacement:** To enrich the verbalized narratives, we enhance non-critical terms with contextually suitable synonyms, drawing on a comprehensive lexical resource (Miller, 1995). This process selectively varies language (up to five substitutions per instance) to improve the diversity of minority class descriptions, ensuring the meaning remains intact and supports robust classification performance.

To maintain data quality, we implement a semantic validation procedure where a random subset of augmented examples (10%) is manually inspected to verify that: (1) class-determining features remain unaltered, (2) causal relationships between features are preserved, and (3) linguistic coherence is maintained. Samples that violate these criteria are discarded (Fernández et al., 2018). This approach ensures that augmented instances remain valid representatives of their respective classes while introducing sufficient linguistic diversity to improve model generalization.

3.3 Parameter-Efficient Fine-Tuning

Finally, we fine-tune a small pre-trained encoder-only language model (under 100M parameters) on the augmented verbalized data to predict the original class labels. This step uses parameter-efficient techniques to adapt this compact model to safety-critical tabular classification tasks while maintaining computational feasibility. This approach contrasts with our verbalization step, which leverages a large generative language model in a zero-shot manner. For classification, we specifically choose an encoder-only architecture that processes input sequences efficiently and produces class predictions directly, making it suitable for deployment in resource-constrained environments.

For our classification model, we select DistilBERT (Sanh et al., 2020), a compact encoder-only language model with only 66 million parameters—significantly smaller than models used in comparable approaches such as TabLLM (11B) (Hegselmann et al., 2023) and TAPAS (340M) (Herzig et al., 2020). This choice is motivated by three factors: (1) encoder-only architectures are well-suited for classification tasks where the output is

a class label rather than generated text; (2) DistilBERT balances computational efficiency and performance, ideal for resource-constrained deployments in safety-critical domains; and (3) through knowledge distillation, DistilBERT retains 97% of BERT’s language understanding capabilities (Sanh et al., 2020) while requiring just 40% of the parameters.

To further enhance efficiency, we implement Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023), which combines 4-bit quantization with low-rank adaptation to reduce memory requirements while maintaining performance. The fine-tuning process consists of the following components:

- **Model architecture:** We use DistilBERT with a sequence classification head that leverages the [CLS] token representation for prediction. The model is initialized with pre-trained weights to leverage transfer learning from general language understanding tasks.
- **QLoRA configuration:** We apply 4-bit quantization to the base model parameters and add trainable low-rank adaptation matrices with rank $r = 16$ and $\alpha = 32$. QLoRA is applied to attention modules (q_lin, v_lin) and classification layers (classifier.dense, classifier.out_proj), with dropout rate 0.05.
- **Training configuration:** We use a maximum sequence length of 512 tokens, learning rate of $1e-3$ with linear warmup over 10% of training steps followed by linear decay, batch size of 32, and train for 20 epochs with early stopping based on validation loss. Weight decay of 0.01 is applied for regularization.

This approach enables efficient adaptation of pre-trained language models to tabular data tasks while requiring significantly fewer computational resources than full fine-tuning or methods using larger models. Training on a single Google Colab A100 GPU completes 20 epochs for our largest dataset (Gas Sensor Array Drift, $\sim 13,910$ samples) in under 1 hour, demonstrating the practical deployability of our framework even in resource-constrained environments.

3.4 Baseline Models and Comparative Analysis

To comprehensively evaluate our approach, we implement nine baseline models spanning traditional ML, ensemble methods, and neural architectures:

- **Traditional ML:** k-Nearest Neighbors (k-NN), Decision Trees, SVMs
- **Ensemble methods:** Random Forests, XGBoost
- **Neural architectures:** MLP, one-dimensional (1D) CNN, Transformer encoder

For traditional ML and ensemble methods, we perform rigorous hyperparameter optimization using stratified k-fold cross-validation ($k=5$) with grid search over extensive parameter spaces (detailed in Appendix A.4). For neural architectures, we adopt standard configurations per literature (Grinsztajn et al., 2022), as exhaustive tuning across six datasets is computationally prohibitive with limited gains, especially for smaller datasets such as Glass (214 samples) (Shwartz-Ziv and Armon, 2021). For fair comparison, all baseline models are trained on original tabular data with SMOTE (Chawla et al., 2002) applied before training, ensuring performance differences stem from our verbalization approach rather than imbalance mitigation.

3.5 Ablation Study

To assess the impact of different components in our framework, we conduct an ablation study focusing on three critical aspects:

1. **Verbalization quality:** We compare our structured verbalization approach with simpler feature-value mapping approaches similar to TabLLM (Hegselmann et al., 2023) to isolate the impact of rich textual descriptions.
2. **Few-shot learning:** We evaluate ChatGPT-4o in zero-shot and 5-shot classification settings using instruction fine-tuning on verbalized instances, separate from the DistilBERT pipeline. Instructions are task-specific, e.g., “Classify whether a semiconductor process fails based on sensor data” for SECOM.
3. **Augmentation strategy:** We compare our linguistically-informed augmentation with SMOTE and no augmentation to quantify the contribution of semantic-preserving text augmentation.

This ablation study is conducted on a subset of three datasets with significant minority class imbalance (AI4I, SECOM, Student Performance) to specifically evaluate the framework’s effectiveness for safety-critical minority class detection. The results provide insights into which components contribute most significantly to performance improvements.

3.6 Evaluation Methodology

We evaluate all models using stratified 80/20 train-test splits with independent verbalization of test instances to prevent data leakage. Performance is assessed via accuracy, precision, recall, and F1-score, emphasizing minority class metrics for unbalanced datasets. For binary tasks (AI4I, Mammographic Mass, SECOM), we report class-specific metrics for both classes; for multi-class tasks (Glass, Gas Sensor, Student), we provide macro-averaged metrics alongside performance for least-represented classes.

4 Experiments and Results

This section presents a comprehensive evaluation of our “numbers to narratives” framework for safety-critical classification across six diverse UCI datasets (Dua and Graff, 2017). We implement our approach as described in Section 3, using stratified 80/20 train-test splits with independent verbalization of test instances to prevent data leakage. All training is completed on a single Google Colab A100 GPU, with 20 epochs for our largest dataset (Gas Sensor Array Drift, $\sim 13,910$ samples) requiring under 1 hour—significantly more efficient than comparable approaches such as TabLLM (11B) and TAPAS (340M) (Hegselmann et al., 2023; Herzig et al., 2020). For fair comparison, baseline models are trained with SMOTE (Chawla et al., 2002), ensuring that performance differences can be attributed to our verbalization approach rather than simply to rebalancing techniques. Our evaluation focuses on both overall accuracy and minority class performance, with particular emphasis on the framework’s effectiveness in addressing the “minority class blind spot” in high-consequence domains.

4.1 Overall Classification Performance

Table 2 summarizes the overall accuracy across all nine models and six datasets. Our DistilBERT-based approach achieves the highest accuracy in five out of six datasets: AI4I (99.25%), Glass Identification (83.39%), Student Performance (80.75%), Gas Sensor Array Drift (99.10%), and Mammographic Mass (91.83%). These results represent improvements of 1.50 to 22.43 percentage points over the best baseline models.

Particularly notable is the substantial improvement on the Gas Sensor Array Drift dataset, where our approach achieves 99.10% accuracy compared to the best baseline (MLP: 76.67%), representing a

Table 2: Overall Accuracy (%) Across Models and Datasets. **Best** performance for each dataset is highlighted in bold, and the second best is underlined. DT: Decision Tree, RF: Random Forest, XGB: XGBoost, TF: Transformer

Dataset	k-NN	DT	RF	SVM	XGB	MLP	CNN	TF	DistilBERT
AI4I	94.30	95.70	96.55	93.55	<u>97.75</u>	96.70	91.40	91.30	99.25
Glass	76.74	72.09	<u>79.07</u>	53.49	62.79	72.09	55.81	72.09	83.39
Student	56.96	58.23	68.35	64.56	58.23	<u>72.15</u>	65.82	<u>72.15</u>	80.75
Gas	66.31	50.98	62.68	75.41	74.46	<u>76.67</u>	74.50	72.58	99.10
Mammographic	81.35	79.79	82.38	80.83	<u>84.97</u>	81.35	82.38	<u>84.97</u>	91.83
SECOM	62.42	86.94	<u>93.31</u>	93.63	84.39	90.13	87.26	89.17	67.09

Table 3: Minority Class F1-Scores (%) for Unbalanced Datasets. The **best** performance for each dataset is highlighted in bold, and the second best is underlined.

Dataset	k-NN	DT	RF	SVM	XGB	MLP	CNN	TF	DistilBERT	F1 Gain vs. 2nd Best	Class Ratio
AI4I	47.22	54.26	59.65	47.35	<u>71.34</u>	62.07	41.89	41.22	78.76	7.42	~24:1
Student	34.71	43.92	38.55	38.89	34.19	<u>53.75</u>	39.98	51.35	65.87	12.12	~3:1 to 5:1
SECOM	18.77	19.85	0.00	0.09	<u>30.99</u>	16.77	19.66	23.48	32.00	1.01	~14:1

22.43 percentage point improvement. This exceptional performance suggests that our verbalization approach is particularly effective for datasets with complex feature interactions that can be meaningfully captured through natural language descriptions.

In the SECOM (semiconductor manufacturing) dataset, our approach achieves 67.09% accuracy, underperforming compared to SVM (93.63%) and Random Forest (93.31%), with a 26.54 percentage point gap. This reflects a deliberate trade-off prioritizing minority class detection in this highly imbalanced dataset (14:1). While these traditional models achieve impressive overall accuracy, their near-zero minority class F1-scores (0.09% and 0.00%) reveal they essentially ignore critical failure cases. In contrast, our approach attains a meaningful 32.00% F1-score for the minority class—a crucial capability for safety-critical applications, as detailed in the next section.

4.2 Safety-Critical Minority Class Performance

For safety-critical applications, minority class performance is of paramount importance, as it directly impacts the reliability of detecting rare but consequential events, particularly in datasets with significant class imbalances. Table 3 presents minority class F1-scores for three unbalanced datasets with notable safety implications: AI4I (machine failure prediction) with a ~49:1 imbalance (98% no failure, 2% failure), Student Performance (at-risk student identification) with a ~3:1 to 5:1 imbalance (77–85% pass, 15–23% fail), and SECOM (semiconductor manufacturing quality) with a 14:1 imbalance (93.4% pass, 6.6% fail).

These extreme disparities underscore the challenge of minority class detection, where traditional models often fail, making our framework’s F1-score improvements—enabled by structured verbalization and augmentation—especially significant for ensuring robust identification of critical anomalies.

Our approach significantly outperforms baseline models in minority class detection across three unbalanced datasets with critical implications: AI4I (machine failure prediction, 49:1 imbalance), Student Performance (at-risk identification, 3:1 to 5:1 imbalance), and SECOM (semiconductor quality, 14:1 imbalance). For AI4I, DistilBERT achieves a 78.76% F1-score (+7.42 points over XGBoost’s 71.34%), vital for predictive maintenance. In Student Performance, it reaches 65.87% (+12.12 points over MLP’s 53.75%), enhancing at-risk student detection through contextual narratives. For SECOM, despite the extreme imbalance, it attains 32.00% (+1.01 points over XGBoost’s 30.99%), outperforming models like Random Forest (0.00%) and SVM (0.09%), demonstrating resilience in detecting rare failures.

This performance reflects an intentional trade-off, with SECOM’s 32.00% F1-score accompanied by a 26.22% overall accuracy drop (67.09% vs. 93.31%), prioritizing minority detection in safety-critical contexts. The substantial gains in AI4I and Student, despite their respective imbalances, underscore the framework’s effectiveness, while SECOM’s challenge highlights opportunities for further imbalance mitigation. This aligns with domains where false negatives (e.g., undetected failures) carry higher costs than false positives, affirming the approach’s practical value.

Table 4: F1-Score Comparison of Best ML, Best DL, DistilBERT with Verbalization, and ChatGPT-4o

Dataset	Best ML	Best DL	Distil-BERT	Zero-Shot	5-Shot
AI4I	85.00	80.00	87.50	49.14	46.39
Glass	81.33	71.17	76.83	47.50	55.67
Student	55.50	67.00	81.50	55.12	66.33
Gas	76.40	76.67	99.00	65.37	72.13
Mammographic	85.00	85.00	87.50	45.00	53.33
SECOM	61.00	58.50	55.50	44.88	58.27

4.3 Component Analysis through Ablation Studies

Table 4 compares the best ML, best DL, verbalization-based DistilBERT, and ChatGPT-4o (zero-shot, 5-shot) approaches, revealing key insights. DistilBERT outperforms ML and DL in four of six datasets (e.g., AI4I: 87.50% vs. 85.00% ML, +2.50–23.00 points), with exceptions in Glass (Random Forest: 81.33%) and SECOM due to its 14:1 imbalance. ChatGPT-4o lags significantly (zero-shot: -33.49%, 5-shot: -25.41% on average), underscoring fine-tuning’s value. While 5-shot improves over zero-shot (e.g., Student: 66.33% vs. 55.12%), it underperforms on AI4I (46.39% vs. 49.14%), suggesting potential interference from subtle feature interactions. Experiments show rich verbalization outperforms TabLLM-style mappings (Hegselmann et al., 2023) by 7.23 F1 points, while our augmentation exceeds SMOTE by 5.45 points (details in Appendix A.3).

4.4 Discussion

Our evaluation demonstrates that the “numbers to narratives” framework significantly enhances safety-critical classification, particularly minority class detection. The approach achieves consistent improvements across datasets with varying characteristics, outperforming traditional methods in both overall accuracy (e.g., AI4I: 99.25%, Student: 80.75%) and minority class metrics (AI4I: 78.76% F1, Student: 65.87% F1) in five of six cases.

Ablation studies quantify the contribution of each component: (1) structured verbalization provides a 7.23 percentage point improvement over simple feature-value mappings like TabLLM (Hegselmann et al., 2023); (2) linguistically-informed augmentation outperforms SMOTE by 5.45 percentage points while preserving causal dependencies; and (3) parameter-efficient fine-tuning significantly outperforms both zero-shot (-33.49 points) and 5-shot (-25.41 points) classification with ChatGPT-4o.

The exceptional performance on Gas Sensor Ar-

ray Drift (99.10% accuracy, +22.43 points over MLP) highlights our approach’s effectiveness for datasets with complex feature interactions that natural language can effectively represent. Conversely, the SECOM dataset’s challenges reveal important limitations when facing extreme imbalance (14:1). Despite achieving a small improvement in minority class detection (+1.01 points, 32.00% F1), the substantial accuracy trade-off (-26.22 points, 67.09%) reflects a deliberate focus on rare event detection.

These comprehensive results address all three research questions from Section 1. Our findings confirm that fine-tuned LMs with verbalized inputs outperform conventional models across diverse dataset characteristics (RQ1), linguistically-informed augmentation significantly improves minority class performance compared to methods like SMOTE (RQ2), and rich contextual verbalization substantially outperforms simple feature-value mappings (RQ3).

These findings have profound implications for domains where reliable detection of rare but consequential events is critical. By bridging the gap between tabular data and natural language, our framework enables language models to apply their pre-trained knowledge to safety-critical classification tasks, effectively addressing the “minority class blind spot” that limits traditional approaches.

5 Conclusion

The “numbers to narratives” framework introduced in this paper transforms tabular data into contextually rich descriptions for improved safety-critical classification. By leveraging language models’ pre-trained knowledge, our approach addresses the “minority class blind spot” in traditional methods while offering dual advantages: enhanced minority class detection and significant computational efficiency. Using a compact 66M-parameter language model with parameter-efficient fine-tuning, our approach achieves superior results with just a fraction of the computational resources required by comparable methods, enabling practical deployment even on single-GPU environments.

Future work should focus on domain-specific verbalization for technical fields with abstract features, advanced augmentation techniques for extreme imbalance, and further optimizing efficiency for resource-constrained environments—extensions that would enhance applicability across diverse safety-critical domains.

6 Limitations

Our framework improves safety-critical classification but faces several notable limitations. First, extreme class imbalance poses significant challenges, as evidenced by SECOM’s 14:1 ratio where DistilBERT achieves 67.09% accuracy and 32.00% minority F1-score—a 26.22% accuracy drop from Random Forest’s 93.31% despite a small +1.01 point improvement in minority detection over XGBoost (30.99%). This trade-off suggests the need for domain-specific augmentation strategies tailored to high-dimensional sensor data (Chawla et al., 2002).

Second, our approach risks overfitting on smaller datasets such as Glass Identification (214 samples), where DistilBERT’s F1-score (76.83%) trails Random Forest’s (81.33%). Adjusting QLoRA hyperparameters such as rank or dropout could improve generalization (Dettmers et al., 2023), addressing the broader challenge of balancing model capacity against overfitting with limited training data.

Third, our verbalization approach introduces computational overhead compared to traditional ML methods, both during training and inference. While QLoRA significantly reduces resource requirements compared to full fine-tuning, the computational cost remains higher than traditional ML models like XGBoost, potentially limiting applicability in resource-constrained environments or real-time systems where latency is critical.

Fourth, the pipeline’s reliance on ChatGPT-4o for text conversion introduces dependency on a proprietary model, potentially limiting reproducibility. Open-source alternatives such as LLaMA (Touvron et al., 2023) could offer more transparent and customizable verbalization processes.

Finally, our ablation study is limited to zero-shot and 5-shot ChatGPT-4o evaluations, which may not fully capture the potential of few-shot learning with more examples or alternative prompting strategies. Additionally, the exceptional performance on Gas Sensor Array Drift (99.10% accuracy, +22.43% over MLP) requires further validation to ensure generalizability. Future work should explore lightweight LLMs or hybrid approaches combining ML and DL strengths (Xu et al., 2023) to balance performance and efficiency.

References

- Sercan O. Arik and Tomas Pfister. 2020. [Tabnet: Attentive interpretable tabular learning](#). [Preprint](#), arXiv:1908.07442.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. [On the opportunities and risks of foundation models](#). [Preprint](#), arXiv:2108.07258.
- Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2024. [Deep neural networks and tabular data: A survey](#). *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519.
- Leo Breiman. 2001. [Random Forests](#). *Mach. Learn.*, 45(1):5–32.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario

815	Amodei. 2020. Language models are few-shot learners. <u>Advances in Neural Information Processing Systems</u> , 33:1877–1901.	
816		
817		
818	Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. <u>A systematic study of the class imbalance problem in convolutional neural networks</u> . <u>Neural Networks</u> , 106:249–259.	
819		
820		
821		
822	Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. <u>Journal of Artificial Intelligence Research</u> , 16:321–357.	
823		
824		
825		
826	Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. 2003. <u>Smoteboost: Improving prediction of the minority class in boosting</u> . In <u>Knowledge Discovery in Databases: PKDD 2003</u> , volume 2838 of <u>Lecture Notes in Computer Science</u> , Berlin, Heidelberg. Springer.	
827		
828		
829		
830		
831		
832	Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In <u>Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining</u> , pages 785–794.	
833		
834		
835		
836		
837	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized large language models. <u>arXiv:2305.14314</u> .	
838		
839		
840		
841	Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml .	
842		
843		
844	Charles Elkan. 2001. The foundations of cost-sensitive learning. In <u>Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI’01</u> , page 973–978, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.	
845		
846		
847		
848		
849	Alberto Fernández, Salvador García, Mikel Galar, Rocío C. Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. <u>Learning from Imbalanced Data Sets</u> . Springer.	
850		
851		
852		
853	Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. <u>Deep Learning</u> . MIT Press.	
854		
855	Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? In <u>Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22</u> , Red Hook, NY, USA. Curran Associates Inc.	
856		
857		
858		
859		
860		
861	Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In <u>2008 IEEE International Joint Conference on Neural Networks</u> , pages 1322–1328.	
862		
863		
864		
865		
866	Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. <u>IEEE Transactions on Knowledge and Data Engineering</u> , 21(9):1263–1284.	
867		
868		
	M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. <u>Support vector machines</u> . <u>IEEE Intelligent Systems and their Applications</u> , 13(4):18–28.	869 870 871 872
	Stefan Hegerselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sonntag. 2023. <u>Tabllm: Few-shot classification of tabular data with large language models</u> . <u>Preprint</u> , arXiv:2210.10723.	873 874 875 876 877
	Jonathan Herzig, Pawel K. Nowak, Thomas Müller, Francesco Piccinno, and Julian M. Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-trained language models. In <u>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</u> , pages 4320–4333.	878 879 880 881 882 883
	Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. <u>arXiv:2106.09685</u> .	884 885 886 887
	Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. <u>Survey on deep learning with class imbalance</u> . <u>Journal of Big Data</u> , 6:27.	888 889 890
	Sotiris B. Kotsiantis, Ioannis D. Zaharakis, and Panayiotis E. Pintelas. 2006. Machine learning: A review of classification and combining techniques. <u>Artificial Intelligence Review</u> , 26(3):159–190.	891 892 893 894
	Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. <u>Table-gpt: Table fine-tuned gpt for diverse table tasks</u> . <u>Proc. ACM Manag. Data</u> , 2(3).	895 896 897 898 899
	Michael McCann and Adrian Johnston. 2008. SECOM Dataset Analysis for Semiconductor Manufacturing.	900 901
	George A. Miller. 1995. Wordnet: A lexical database for english. <u>Communications of the ACM</u> , 38(11):39–41.	902 903 904
	Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. <u>Rethinking the role of demonstrations: What makes in-context learning work?</u> In <u>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</u> , pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	905 906 907 908 909 910 911 912
	Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. <u>Facebook FAIR’s WMT19 news translation task submission</u> . In <u>Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)</u> , pages 314–319, Florence, Italy. Association for Computational Linguistics.	913 914 915 916 917 918 919
	OpenAI. 2024. ChatGPT-4o: A Multimodal Large Language Model. https://openai.com/research/gpt-4o .	920 921 922

923	Sergei Popov, Stanislav Morozov, and Artem Babenko.	Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui	978
924	2019. Neural oblivious decision ensembles for deep	Tao, and Fu Lee Wang. 2023. Parameter-efficient	979
925	learning on tabular data . Preprint , arXiv:1909.06312.	fine-tuning methods for pretrained language mod-	980
926	Foster Provost and Tom Fawcett. 2013. Data Science	els: A critical review and assessment . Preprint ,	981
927	for Business: What You Need to Know about Data	arXiv:2312.12148.	982
928	Mining and Data-Analytic Thinking . O'Reilly Me-	Jingfeng Yang, Aditya Gupta, Shyam Upadhyay,	983
929	dia.	Luheng He, Rahul Goel, and Shachi Paul. 2022.	984
930	Victor Sanh, Lysandre Debut, Julien Chaumond, and	Tableformer: Robust transformer modeling for table-	985
931	Thomas Wolf. 2020. Distilbert, a distilled version	text encoding . Preprint , arXiv:2203.00274.	986
932	of bert: smaller, faster, cheaper and lighter . Preprint ,	Pengcheng Yin, Graham Neubig, Wen tau Yih, and	987
933	arXiv:1910.01108.	Sebastian Riedel. 2020. Tabert: Pretraining for joint	988
934	Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse,	understanding of textual and tabular data . Preprint ,	989
935	and Amri Napolitano. 2010. Rusboost: A hy-	arXiv:2005.08314.	990
936	brid approach to alleviating class imbalance . IEEE		
937	Transactions on Systems, Man, and Cybernetics -		
938	Part A: Systems and Humans , 40(1):185–197.		
939	Rico Sennrich, Barry Haddow, and Alexandra Birch.		
940	2016. Neural machine translation of rare words with		
941	subword units. In Proceedings of the 54th Annual		
942	Meeting of the Association for Computational		
943	Linguistics , pages 1715–1725.		
944	Ravid Shwartz-Ziv and Amitai Armon. 2021. Tabular		
945	data: Deep learning is not all you need . Preprint ,		
946	arXiv:2106.03253.		
947	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier		
948	Martinet, Marie-Anne Lachaux, Timothée Lacroix,		
949	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal		
950	Azhar, Aurelien Rodriguez, Armand Joulin, Edouard		
951	Grave, and Guillaume Lample. 2023. LLaMA: Open		
952	and Efficient Foundation Language Models . arXiv		
953	preprint . ArXiv:2302.13971 [cs].		
954	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Us-		
955	zoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser,		
956	and Illia Polosukhin. 2017. Attention is all you		
957	need. In Advances in Neural Information Processing		
958	Systems , volume 30, pages 5998–6008.		
959	Alexander Vergara, Shankar Vembu, Tuba Ayhan, Mar-		
960	garet A. Ryan, Margie L. Homer, and Ramón Huerta.		
961	2012. Chemical gas sensor drift compensation us-		
962	ing classifier ensembles. Sensors and Actuators B:		
963	Chemical , 166–167:320–329.		
964	Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi		
965	Fu, Shi Han, and Dongmei Zhang. 2021. Tuta:		
966	Tree-based transformers for generally structured ta-		
967	ble pre-training . In Proceedings of the 27th ACM		
968	SIGKDD Conference on Knowledge Discovery &		
969	Data Mining , KDD '21, page 1780–1790, New York,		
970	NY, USA. Association for Computing Machinery.		
971	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten		
972	Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le,		
973	and Denny Zhou. 2022. Chain-of-thought prompt-		
974	ing elicits reasoning in large language models. In		
975	Proceedings of the 36th International Conference on		
976	Neural Information Processing Systems , NIPS '22,		
977	Red Hook, NY, USA. Curran Associates Inc.		

A Supplementary Results and Specifications

This appendix provides detailed performance metrics and configurations for the experiments presented in the main paper. It includes comprehensive tables summarizing overall performance (Section A.1), minority class performance (Section A.2), zero-shot and few-shot performance of GPT-4o (Section A.3), the system prompt used for tabular-to-text conversion (Section A.4.1), and hyperparameters for baseline models (Section A.4.2), dataset-specific configurations (Section A.4.3), and DistilBERT QLoRA (Section A.4.4).

A.1 Detailed Overall Performance

Table 5 presents the overall performance metrics across all evaluated datasets, including accuracy, precision, recall, and F1-score for each model type and dataset.

Table 5: Overall Performance Across Datasets

Model Type	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AI4I Dataset					
Conventional ML	k-NN	94.30	66.50	85.00	72.00
	Decision Tree	95.70	70.50	85.50	76.00
	Random Forest	96.55	73.50	86.00	79.00
	SVM	93.55	66.00	89.50	72.00
Deep Learning	XGBoost	97.75	81.00	90.00	85.00
	MLP	96.70	75.00	88.00	80.00
	1D CNN	91.40	62.50	91.00	68.50
	Transformer	91.30	62.50	90.50	68.00
Large Language Model	DistilBERT	99.25	93.50	85.00	87.50
Glass Identification Dataset					
Conventional ML	k-NN	76.74	73.83	82.33	76.17
	Decision Tree	72.09	76.17	84.33	76.33
	Random Forest	79.07	80.00	89.50	81.33
	SVM	53.49	61.83	72.83	60.17
Deep Learning	XGBoost	62.79	64.67	77.33	67.17
	MLP	72.09	69.83	73.17	71.17
	1D CNN	55.81	67.67	60.17	54.83
	Transformer	72.09	67.00	79.00	69.83
Large Language Model	DistilBERT	83.39	79.17	73.67	76.83
Student Performance Dataset					
Conventional ML	k-NN	56.96	53.50	63.50	46.50
	Decision Tree	58.23	55.50	57.50	55.50
	Random Forest	68.35	47.50	50.00	47.50
	SVM	64.56	97.00	52.50	53.50
Deep Learning	XGBoost	58.23	52.50	52.50	52.00
	MLP	72.15	68.00	66.50	67.00
	1D CNN	65.82	59.50	58.00	58.00
	Transformer	72.15	68.00	65.50	66.00
Large Language Model	DistilBERT	80.75	78.50	83.00	81.50
Gas Sensor Array Drift Dataset					
Conventional ML	k-NN	66.31	66.83	66.00	65.67
	Decision Tree	50.98	51.50	51.00	51.17
	Random Forest	62.68	64.22	62.71	62.74
	SVM	75.41	79.78	75.51	76.40
Deep Learning	XGBoost	74.46	77.42	75.33	74.87
	MLP	76.67	77.00	76.67	76.67
	1D CNN	74.50	78.00	74.33	74.33
	Transformer	72.58	74.50	72.67	74.33
Large Language Model	DistilBERT	99.10	99.33	99.00	99.00
Mammographic Mass Dataset					
Conventional ML	k-NN	81.35	81.50	81.50	81.50
	Decision Tree	79.79	80.50	80.50	80.00
	Random Forest	82.38	83.00	83.00	82.00
	SVM	80.83	81.50	81.50	81.00
Deep Learning	XGBoost	84.97	85.00	85.50	85.00
	MLP	81.35	84.00	81.50	81.50
	1D CNN	82.38	84.00	83.00	82.50
	Transformer	84.97	85.00	85.50	85.00
Large Language Model	DistilBERT	91.83	89.50	86.00	87.50
SECOM Dataset					
Conventional ML	k-NN	62.42	53.50	63.50	46.50
	Decision Tree	86.94	56.50	58.50	56.50
	Random Forest	93.31	47.50	50.00	48.50
	SVM	93.63	97.00	52.50	53.00

Table 6: Minority Class Performance

Model Type	Model	Precision (%)	Recall (%)	F1-Score (%)
AI4I Dataset (Class: Machine Failure)				
Conventional ML	k-NN	34.46	75.00	47.22
	Decision Tree	42.50	75.00	54.26
	Random Forest	49.51	75.00	59.65
	SVM	32.77	85.29	47.35
	XGBoost	62.92	82.35	71.34
Deep Learning	MLP	50.94	79.41	62.07
	1D CNN	27.19	91.18	41.89
Transformer	Transformer	26.75	89.71	41.22
Large Language Model	DistilBERT	85.19	72.50	78.76
Student Performance (Class: Fail)				
Conventional ML	k-NN	35.67	35.27	34.71
	Decision Tree	38.75	49.75	43.92
	Random Forest	52.41	30.34	38.55
	SVM	44.75	34.19	38.89
	XGBoost	35.50	34.85	34.19
Deep Learning	MLP	58.51	49.55	53.75
	1D CNN	46.66	34.87	39.98
Transformer	Transformer	59.55	46.24	51.35
Large Language Model	DistilBERT	66.75	58.55	65.87
SECOM (Class: Semiconductor Failure)				
Conventional ML	k-NN	10.55	66.68	18.77
	Decision Tree	16.66	24.45	19.85
	Random Forest	0.00	0.00	0.00
	SVM	1.00	0.05	0.09
	XGBoost	22.00	52.38	30.99
Deep Learning	MLP	18.35	14.34	16.77
	1D CNN	16.76	23.87	19.66
Transformer	Transformer	22.34	23.75	23.48
Large Language Model	DistilBERT	50.00	23.00	32.00

Table 5 – continued from previous page

Model Type	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Deep Learning	XGBoost	84.39	59.00	69.50	61.00
	MLP	90.13	56.50	55.00	55.50
	1D CNN	87.26	55.50	58.00	56.50
Transformer	Transformer	89.17	58.50	59.00	58.50
Large Language Model	DistilBERT	67.09	60.00	56.50	55.50

A.2 Detailed Minority Class Performance

Table 6 provides performance metrics for the minority class across the AI4I, Student Performance, and SECOM datasets, highlighting the effectiveness of our approach in handling class imbalance. Metrics include precision, recall, and F1-score for the minority class.

A.3 Detailed Zero-Shot and Few-Shot GPT-4o Performance

Table 7 details the performance of GPT-4o in zero-shot and few-shot (5-shot) settings across all datasets. Metrics include accuracy, precision, recall, and F1-score for overall and minority class performance.

Table 7: Detailed Zero-Shot and Few-Shot GPT-4o Performance

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AI4I Dataset					
AI4I	Zero-Shot	96.60	48.30	50.00	49.14

Continued on next page

Table 7 – continued from previous page

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
	Zero-Shot (Minority)	–	0.00	0.00	0.00
	5-Shot	65.85	53.38	73.81	46.39
	5-Shot (Minority)	–	26.54	82.35	24.38
Student Performance Dataset					
Student Performance	Zero-Shot	60.76	55.19	55.08	55.12
	Zero-Shot (Minority)	–	40.00	38.46	39.22
	5-Shot	70.32	65.58	68.20	66.33
	5-Shot (Minority)	–	58.22	52.12	50.65
SECOM Dataset					
SECOM	Zero-Shot	83.77	45.38	46.44	44.88
	Zero-Shot (Minority)	–	12.45	26.55	21.54
	5-Shot	88.36	52.33	56.23	58.27
	5-Shot (Minority)	–	33.87	26.65	25.88
Glass Identification Dataset					
Glass Identification	Zero-Shot	62.79	46.67	48.33	47.50
	Zero-Shot (Minority)	–	30.00	28.57	29.27
	5-Shot	67.44	54.17	57.14	55.67
	5-Shot (Minority)	–	40.00	42.86	41.38
Gas Sensor Array Drift Dataset					
Gas Sensor Array Drift	Zero-Shot	70.45	64.29	66.67	65.37
	Zero-Shot (Minority)	–	50.00	48.00	49.00
	5-Shot	75.63	70.83	73.33	72.13
	5-Shot (Minority)	–	55.56	60.00	57.69
Mammographic Mass Dataset					
Mammographic Mass	Zero-Shot	60.83	44.44	45.45	45.00
	Zero-Shot (Minority)	–	33.33	31.25	32.26
	5-Shot	65.62	52.38	54.55	53.33
	5-Shot (Minority)	–	41.67	45.45	43.48

A.4 Experimental Settings

This section details the experimental configurations, including the system prompt used for tabular-to-text conversion and the hyperparameters for baseline models and DistilBERT.

A.4.1 System Prompt

The following system prompt was used to configure models for tabular-to-text conversion, ensuring consistent instruction for predictive maintenance and classification tasks across datasets:

System Prompt – Tabular-to-Text Conversion

You are a data annotation assistant specialized in transforming structured tabular data into instruction-tuned text for language model fine-tuning. For each dataset, summarize the following in a concise format:

- **Dataset Overview:** Provide the name, domain, purpose, size, and context in one line. [e.g., “SECOM, Semiconductor Manufacturing, Predict process failures, 1567 samples, Sensor data with noise”]
- **Feature Details:** List included features with name, description, type, and any relevant mappings; mention excluded features if any.
- **Target Variable:** Specify the target column, its type (e.g., binary, multiclass, regression), and label mappings (e.g., 0=Pass, 1=Fail).

A.4.2 Baseline Model Hyperparameters

Table 8 lists the hyperparameter search spaces for the baseline models (k-NN, Decision Tree, Random Forest, SVM, XGBoost, MLP, 1D CNN, and Transformer) used in our experiments. These settings were

Table 8: Baseline Model Hyperparameters

Model	Hyperparameters
k-NN	$n_neighbors \in \{3, 5, 7, 9\}$; $weights \in \{uniform, distance\}$; $metric \in \{euclidean, manhattan\}$
Decision Tree	$max_depth \in \{None, 10, 20, 30\}$; $min_samples_split \in \{2, 5, 10\}$; $min_samples_leaf \in \{1, 2, 4\}$
Random Forest	$n_estimators \in \{50, 100, 200\}$; $max_depth \in \{None, 10, 20\}$; $min_samples_split \in \{2, 5\}$; $min_samples_leaf \in \{1, 2\}$
SVM	$C \in \{0.1, 1, 10\}$; $kernel \in \{linear, rbf\}$; $class_weight \in \{balanced, None\}$
XGBoost	$n_estimators \in \{50, 100, 200\}$; $max_depth \in \{3, 4, 5\}$; $learning_rate \in \{0.01, 0.1, 0.2\}$; $subsample \in \{0.8, 1\}$; $colsample_bytree \in \{0.8, 1\}$
MLP	3 hidden layers: 64, 32, 16 units; ReLU activation; sigmoid/softmax output; Adam optimizer; 50 epochs; batch size 32; learning rate 0.001
1D CNN	Conv1D: 64 filters, kernel size 5; dense layers: 64, 32 units; dropout 0.5; sigmoid/softmax output; 50 epochs; batch size 32
Transformer	3 encoder blocks; 8 attention heads; embedding dimension 256; FFN dimension 512; dense layer: 32 units, ReLU activation; sigmoid/softmax output; 50 epochs; batch size 32

optimized to ensure a fair comparison with our proposed approach.

A.4.3 Dataset-Specific Hyperparameters

Table 9 specifies the optimal hyperparameters selected for the ML baseline models (k-NN, Decision Tree, Random Forest, SVM, and XGBoost) for each dataset, ensuring tailored configurations for optimal performance.

A.4.4 DistilBERT QLoRA Hyperparameters

Table 10 lists the QLoRA hyperparameters used for fine-tuning DistilBERT, optimized for efficient adaptation to the classification tasks.

Table 9: Dataset-Specific Hyperparameters for ML Baselines and DistilBERT

Dataset	k-NN	Decision Tree	Random Forest	SVM	XGBoost
AI4I	$n_neighbors = 3$, $weights : distance$, $metric : euclidean$	$max_depth : None$, $min_samples_split = 2$, $min_samples_leaf = 1$	$n_estimators = 50$, $max_depth = 20$, $min_samples_split = 2$, $min_samples_leaf = 1$	$C = 10$, $kernel : rbf$, $class_weight : balanced$	$n_estimators = 200$, $max_depth = 5$, $learning_rate = 0.2$, $subsample = 0.8$, $colsample_bytree = 1$
Glass Identification	$n_neighbors = 3$, $weights : uniform$, $metric : euclidean$	$max_depth : None$, $min_samples_split = 2$, $min_samples_leaf = 1$	$n_estimators = 50$, $max_depth : None$, $min_samples_split = 2$, $min_samples_leaf = 1$	$C = 0.1$, $kernel : linear$, $class_weight : balanced$	$n_estimators = 50$, $max_depth = 3$, $learning_rate = 0.01$, $subsample = 0.8$, $colsample_bytree = 0.8$
Student Performance	$n_neighbors = 9$, $weights : distance$, $metric : manhattan$	$max_depth = 10$, $min_samples_split = 2$, $min_samples_leaf = 2$	$n_estimators = 100$, $max_depth : None$, $min_samples_split = 2$, $min_samples_leaf = 2$	$C = 10$, $kernel : rbf$, $class_weight : balanced$	$n_estimators = 50$, $max_depth = 4$, $learning_rate = 0.1$, $subsample = 1$, $colsample_bytree = 0.8$
Gas Sensor Array Drift	$n_neighbors = 3$, $weights : uniform$, $metric : euclidean$	$max_depth : None$, $min_samples_split = 2$, $min_samples_leaf = 1$	$n_estimators : None$, $max_depth = 1$, $min_samples_split = 2$, $min_samples_leaf = 50$	$C = 0.1$, $kernel : linear$, $class_weight : balanced$	$n_estimators = 50$, $max_depth = 3$, $learning_rate = 0.1$, $subsample = 0.8$, $colsample_bytree = 0.8$
Mammographic Mass	$n_neighbors = 9$, $weights : uniform$, $metric : manhattan$	$max_depth : None$, $min_samples_split = 10$, $min_samples_leaf = 4$	$n_estimators = 50$, $max_depth = 10$, $min_samples_split = 2$, $min_samples_leaf = 2$	$C = 10$, $kernel : linear$, $class_weight : balanced$	$n_estimators = 100$, $max_depth = 4$, $learning_rate = 0.01$, $subsample = 0.8$, $colsample_bytree = 0.8$
SECOM	$n_neighbors = 3$, $weights : uniform$, $metric : manhattan$	$max_depth : None$, $min_samples_split = 5$, $min_samples_leaf = 1$	$n_estimators = 100$, $max_depth : None$, $min_samples_split = 2$, $min_samples_leaf = 1$	$C = 10$, $kernel : rbf$, $class_weight : balanced$	$n_estimators = 50$, $max_depth = 3$, $learning_rate = 0.01$, $subsample = 0.8$, $colsample_bytree = 0.8$

Table 10: DistilBERT QLoRA Hyperparameters

Parameter	Value
Rank (r)	16
Alpha (α)	32
Quantization Bits	4
Dropout	0.05
Target Layers	['q_lin', 'v_lin', 'classifier.dense', 'classifier.out_proj']
Max Input Length	512
Learning Rate	1e-3
Batch Size	32
Epochs	20
Weight Decay	0.01