

Text Augmentation and Machine Learning for Classifying Internal Quality-Audit Findings

Husam Alghamdi

Industrial Engineering Department

College of Engineering

King Saud University, Riyadh, Saudi Arabia

445107779@student.ksu.edu.sa

Abstract

This study evaluates whether text augmentation improves classification of internal quality-audit findings into ISO 9001:2015 clause categories under limited and imbalanced labeled data. The dataset comprised 191 cleaned findings across seven classes. A domain-aware NLP pipeline was combined with TF-IDF features and three classical classifiers: Logistic Regression, Support Vector Machine (SVM), and Random Forest. Four data settings were compared: no augmentation, SynonymAug, BackTranslation, and ContextualAug. For augmented settings, only training folds were expanded to 2x, 3x, and 5x; test folds were left unchanged. Performance was evaluated with repeated stratified group 5-fold cross-validation using Accuracy, Macro Precision, Macro Recall, and Macro F1. All augmentation methods improved on the no-augmentation baseline, although gains were modest. The strongest result within this experiment was BackTranslation + SVM at 5x, which achieved Accuracy = 0.742, Macro Precision = 0.719, Macro Recall = 0.679, and Macro F1 = 0.680. The findings provide comparative evidence that controlled augmentation can improve clause-level classification of short audit text under constrained-data conditions, but the benefit depends on both augmentation strategy and classifier. The study also outlines a practical dashboard use case for audit triage and hotspot monitoring.

Keywords

internal quality audit, text augmentation, ISO 9001:2015, text classification, machine learning, audit findings

1. Introduction

Internal quality-audit findings are mainly recorded as short narrative observations. They capture evidence of compliance gaps, missing records, weak implementation, and process breakdowns, yet they are still sorted and interpreted largely through manual reading. Linking each finding to the relevant ISO 9001:2015 clause is therefore slow, labor-intensive, and dependent on auditor judgment.

The problem is intensified by the nature of audit text. Findings are short, technical, and rich in abbreviations, document codes, and language of absence or partial implementation. Reliable labels are also costly because clause assignment requires both audit expertise and contextual interpretation. In practice, the bottleneck is rarely the amount of raw text; it is the scarcity of clean and trustworthy labeled examples.

Text augmentation offers a practical response to this constraint. By generating additional labeled variants from existing findings, it can enlarge the effective training set and introduce lexical or structural variation without the same annotation burden. Its value in this setting is therefore not only linguistic variation, but also added training coverage. Recent literature treats augmentation as useful in low-resource NLP, but not as a universal remedy. Reported gains depend on the task, the classifier, and the transformation itself (Feng et al., 2021; Chen et al., 2023). Synonym-based methods are simple and inexpensive, yet they can distort technical meaning; contextual augmentation aims to reduce that risk by conditioning substitutions on sentence context (Wei and Zou, 2019; Kobayashi, 2018). Back-translation

often preserves sentence-level meaning more naturally and has shown strong results in small-data classification, although its gains tend to narrow as real training data increases (Shleifer, 2019; Bayer et al., 2023).

These trade-offs are especially important for internal audit findings, where a few terms may carry much of the class signal. Small wording changes can weaken the link between a finding and its ISO clause, while class imbalance further complicates learning.

That sensitivity changes what counts as a useful augmentation in this setting. In open-domain benchmarks, modest lexical variation may be enough to enrich the training set. In audit findings, however, a single term can signal a missing record, incomplete implementation, or the absence of formal review, and such cues often carry much of the clause signal used for classification. The question is therefore not only whether augmentation enlarges the dataset, but whether it preserves the phrasing on which the label depends. This makes internal audit text a demanding test case for standard augmentation methods.

A second practical issue concerns augmentation size. Expanding a training fold to two, three, or five times its original size can improve coverage and reduce dependence on repeated expressions, yet larger synthetic sets may also introduce redundancy and progressively weaker examples. In small industrial datasets, where authentic labeled findings are scarce and classes are already unevenly distributed, this trade-off becomes practical rather than theoretical. The present study therefore treats augmentation size as a controlled experimental factor rather than assuming that more generated text will always help.

This paper compares four data settings - no augmentation, SynonymAug, BackTranslation, and ContextualAug - within one workflow that combines domain-aware preprocessing, TF-IDF features, and three classical classifiers (Logistic Regression, SVM, and Random Forest). By testing original, 2x, 3x, and 5x training sizes under the same validation design, the study identifies a practical configuration for clause-level classification of small, imbalanced audit text.

1.1 Objectives

The study evaluates whether controlled text augmentation improves classification of internal quality-audit findings into seven ISO 9001:2015 clause categories when labeled data are limited and imbalanced. Specifically, it develops a domain-aware NLP workflow, compares no augmentation with SynonymAug, BackTranslation, and ContextualAug, tests 2x, 3x, and 5x training expansion across Logistic Regression, SVM, and Random Forest, and considers how the best model can support dashboard-based audit triage. Taken together, the design treats augmentation as a combined modeling choice rather than as an isolated preprocessing step by examining transformation type, synthetic-data volume, and classifier response within one evaluation protocol.

The study was guided by three research questions:

RQ1. Which augmentation method gives the best performance for classifying internal quality-audit findings?

RQ2. How does expanding the augmented training set to 2x, 3x, and 5x affect performance?

RQ3. Which augmentation-method-classifier combination is most reliable for ISO 9001:2015 clause classification?

The contribution is both methodological and practical: a controlled comparison framework for small audit-text data and a deployment concept for management use.

2. Literature Review

Text-classification research has advanced quickly, but much of that progress relies on large curated datasets. Industrial texts such as audit findings, inspection notes, and nonconformance statements present a different setting: they are short, technical, unevenly distributed across classes, and expensive to annotate. For that reason, recent surveys treat augmentation as part of limited-data learning rather than as a routine preprocessing step, and they caution that its value is task- and model-dependent (Feng et al., 2021; Chen et al., 2023).

Across the main augmentation families, the central trade-off is between simplicity and meaning preservation. Synonym-based methods are light, practical, and easy to combine with conventional TF-IDF pipelines, which explains why they remain a common baseline in small-data studies (Wei and Zou, 2019). Their limitation is equally clear: a

replacement that appears acceptable at the token level can still shift the function of a technical sentence by moving the emphasis from execution to evidence, from procedure to review, or from omission to delay. For that reason, lexical substitution is best viewed as useful when tightly controlled, not when applied mechanically across sensitive corpora. BackTranslation and contextual augmentation respond to this weakness in different ways. BackTranslation paraphrases the sentence more globally and often preserves communicative intent while varying the surface form enough to broaden coverage (Shleifer, 2019; Kumar et al., 2020). Contextual augmentation instead keeps the sentence frame and changes selected terms according to surrounding context, which can improve grammatical fit and local plausibility (Kobayashi, 2018). Read together, the literature suggests that the difference matters most when a task depends on a small number of meaning-bearing terms. Where class labels are tied to those terms, sentence-level paraphrase can be safer than apparently natural local substitution, especially if the resulting text is later represented through sparse lexical features.

The literature also converges on a second point: augmentation should be evaluated jointly with data volume and model family. Controlled expansion often helps when the original sample is small, but the benefit narrows when synthetic examples become repetitive, weakly informative, or semantically noisy (Bayer et al., 2023; Pellicer et al., 2023). Different classifiers also react differently to the same augmented corpus, so augmentation and model selection should be treated as a combined design choice rather than as separate decisions (Ansari et al., 2021; Kapusta et al., 2024). This matters in technical quality text, where short domain-specific statements, class imbalance, and the need for interpretability all shape what counts as a useful improvement.

Related engineering studies reinforce this point. Work on construction quality problems and defect reports shows that short operational text can be classified effectively, but only when imbalance handling, narrow vocabulary, and interpretability are treated with care (Moreo et al., 2016; Wang et al., 2024). Audit-text studies similarly show that audit language contains enough structure for automated categorization, yet they offer limited direct guidance on how standard augmentation methods behave when labels are scarce, classes overlap, and transparent classical models are preferred for deployment (Vaishampayan et al., 2023). The gap, therefore, is not whether audit text can be classified, but which augmentation strategy, augmentation size, and classical model work best under realistic audit-data constraints.

Viewed as a whole, this literature suggests that effective augmentation in technical text is less about maximizing surface novelty than about preserving decision-relevant wording while broadening the observed expression space. A method can therefore generate fluent alternatives and still be unhelpful if it weakens the lexical cues that carry the class label. This perspective is especially relevant for internal audits, where clause assignment depends on sparse operational cues and where transparent models remain attractive because they make the effect of augmentation easier to interpret.

This study addresses that gap through a controlled comparison of no augmentation, SynonymAug, BackTranslation, and ContextualAug across original, 2x, 3x, and 5x training sizes, using the same preprocessing, feature representation, and validation protocol throughout.

3. Methods

The study used a controlled NLP and machine-learning workflow to classify internal quality-audit findings into ISO 9001:2015 clause categories. The design was intended to isolate the effects of augmentation method, augmented training size, and classifier under a common preprocessing and validation pipeline.

3.1 Research Design

Each finding was treated as a supervised multi-class example and assigned to one of seven ISO 9001:2015 classes: Context of the Organization, Leadership, Planning, Support, Operation, Performance Evaluation, or Improvement. The source data came from a real internal audit management system used in a large engineering consultancy. The raw extract contained 203 records with operational fields such as audit date, reference number, business unit, clause, and finding details. After screening, cleaning, and duplicate removal, 191 findings remained for modeling. For confidentiality, the organization and platform are not named. In routine practice, each finding is assigned to a main ISO clause by the lead auditor and reviewed by the quality head; that operational clause structure was used as the labeling basis in this study.

The decision to retain clause labels assigned in the operational audit workflow reflects the practical setting of the study. These labels were created for real audit closure and follow-up rather than for a benchmark corpus. This makes the task realistic, but it also introduces the kind of boundary ambiguity common in applied classification problems, particularly where findings touch documentation, execution, and review at the same time. The modeling design therefore emphasized controlled comparison and leakage prevention rather than aggressive optimization on a single split.

Accordingly, the dataset was treated as an applied decision-support resource rather than as a benchmark corpus. The goal was not only to maximize a single headline score, but to compare configurations that could later be explained and deployed in a routine audit workflow. This practical orientation influenced model choice, evaluation design, and the decision to keep the feature representation transparent.

The experiment varied three factors: data condition (no augmentation, SynonymAug, BackTranslation, and ContextualAug), augmented training size (original, 2x, 3x, and 5x for augmented settings), and classifier (Logistic Regression, SVM, and Random Forest). Figure 1 summarizes the workflow.

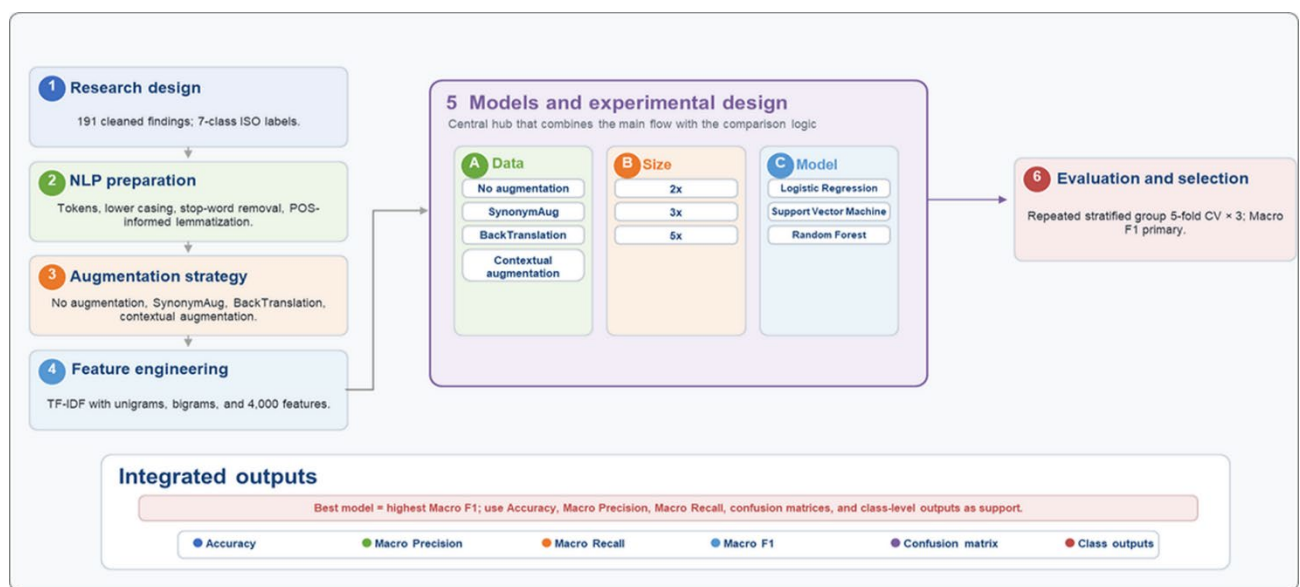


Figure 1. Integrated experimental framework for ISO 9001:2015 clause classification of internal audit findings

3.2 NLP Preparation

Audit findings were tokenized, converted to lower case, stripped of punctuation, filtered for stop words, and then lemmatized using part-of-speech information. Tokens of three characters or fewer were removed from the final corpus. The same pipeline was applied to all records so that later differences could be attributed to augmentation and modeling rather than inconsistent text preparation.

3.3 Data Augmentation Strategy

Four data branches were evaluated: Original_NoAug, SynonymAug, BackTranslation, and ContextualAug. Augmentation was applied only after each train-test split and only to the training fold; the test fold remained unchanged. The archived workflow preserves one retained augmented variant per original finding for each augmentation branch. SynonymAug produced controlled lexical substitutions, BackTranslation produced sentence-level paraphrases, and ContextualAug produced context-aware rewrites.

For each augmented branch, the training fold was expanded to 2x, 3x, and 5x its original size to test the effect of synthetic data volume. Figure 2 shows representative variants from the three augmentation branches. Because the archive records outputs rather than a separate semantic-drift score, fidelity is interpreted through the retained examples and downstream classification results.

The branch structure was chosen to compare three distinct forms of augmentation under the same conditions: lexical substitution, sentence-level paraphrasing, and context-aware rewriting. In all cases, augmentation was treated as training-data expansion rather than as a corpus-wide preprocessing step. This distinction matters in small datasets because applying augmentation before splitting can leak transformed versions of the same finding into train and test partitions. The preserved archive of original and augmented outputs therefore served as a traceable record of how synthetic variants entered each branch, even though it did not include a separate manual drift score for every generated sentence.

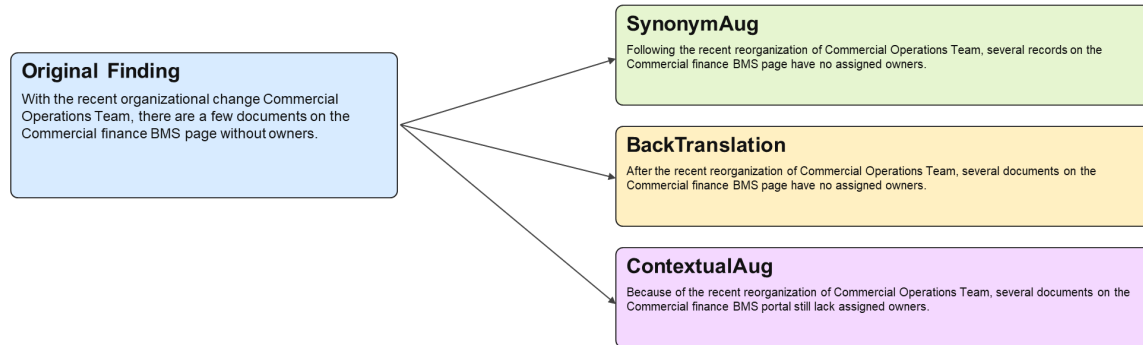


Figure 2. Example augmented variants of one audit finding produced by the three methods.

3.4 Feature Engineering

After preprocessing and, where relevant, augmentation, the text was converted to TF-IDF features using the final filtered lemmatized corpus. The vectorizer used unigrams and bigrams, sublinear term-frequency scaling, and a maximum vocabulary of 4,000 features. It was fitted only on the training portion of each split and then applied to the corresponding test fold.

3.5 Classification Models

Three classifiers were compared: Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). LR provided a strong sparse-text baseline, SVM represented a margin-based model well suited to high-dimensional TF-IDF features, and RF offered a non-linear ensemble comparison. All models used the same feature space.

3.6 Experimental Design

Performance was estimated using repeated stratified group 5-fold cross-validation with three repeats, resulting in 15 evaluation splits. This approach was preferred over a single holdout due to the small and imbalanced dataset. In each split, the train-test partition was created first, and augmentation (if applicable) was applied solely to the training fold until the target size was reached. The TF-IDF vectoriser was fitted only on the training data, and then the classifier was trained and evaluated on the untouched test fold. This validation design was chosen specifically to reduce dependence on a single favorable split and to use the limited labeled data more efficiently than a fixed holdout design, while still preserving class balance and preventing grouped leakage between train and test folds.

The sequence of split first, augment second was fixed across all configurations. This ensured that differences between methods could be attributed to augmentation behaviour and classifier response, rather than artefacts from inconsistent feature fitting or leakage. Given the modest size of the corpus and the rarity of some classes, repeated grouped cross-validation was also crucial in reducing the chance that a single favourable split would unduly influence the reported results result.

3.7 Evaluation Metrics and Best-Model Selection

Performance was assessed using Accuracy, Macro Precision, Macro Recall, and Macro F1. Macro F1 was the primary selection criterion because it weights all seven classes equally and is therefore more suitable than accuracy alone for this imbalanced multi-class task. The remaining metrics were used to check whether the leading model was consistently balanced across classes.

4. Results and Discussion

4.1 Dataset Profile and Class Distribution

After cleaning and duplicate removal, the dataset decreased from 203 to 191 findings. Class distribution remained imbalanced: Operation was largest with 47 findings, followed by Support (42), Planning (40), Performance Evaluation (28), Leadership (14), and Improvement (14), while Context of the Organization contained only six findings. This imbalance makes macro-level metrics more informative than accuracy alone.

4.2 Overall Comparison of Augmentation Methods

Figure 3 compares the best result within each augmentation branch and its Macro-F1 gain over the no-augmentation baseline. The strongest baseline was SVM without augmentation (Accuracy = 0.736, Macro Precision = 0.701, Macro Recall = 0.656, Macro F1 = 0.659). All three augmentation methods improved on this baseline at their best size settings. The strongest result within this experiment was BackTranslation + SVM at 5x, with Accuracy = 0.742, Macro Precision = 0.719, Macro Recall = 0.679, and Macro F1 = 0.680. SynonymAug + SVM at 3x followed closely at Macro F1 = 0.679, while ContextualAug + SVM at 2x reached 0.676.

The gains were modest but consistent at the macro level: SynonymAug improved Macro F1 by +0.020, BackTranslation by +0.021, and ContextualAug by +0.017. This suggests that augmentation improved class balance more than it improved raw accuracy. BackTranslation should therefore be read as the strongest method in this experimental setting, not as a generally superior augmentation strategy (Figure 3).

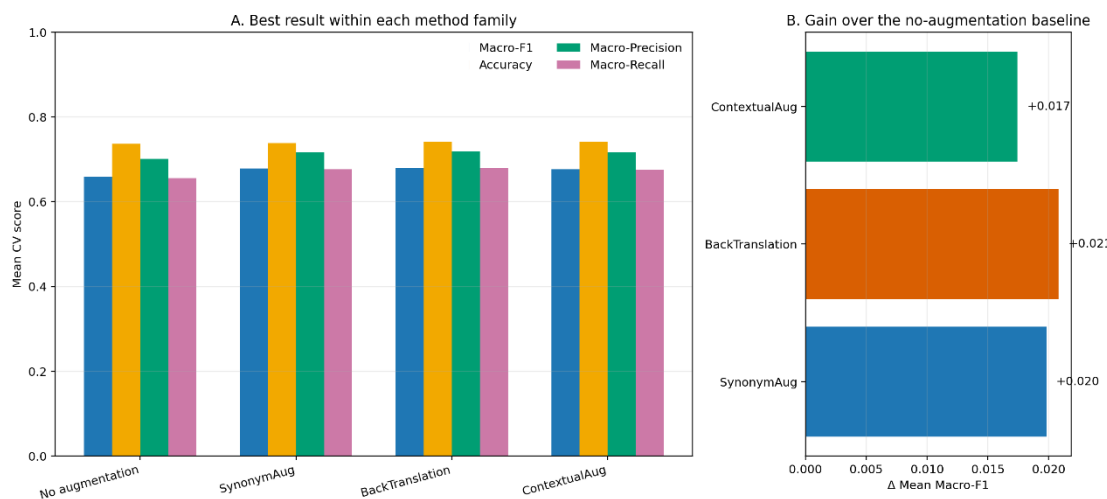


Figure 3. Comparison of Augmentation Methods

4.3 Effect of Augmented Training Size and Classifier

Figure 4 shows that more synthetic data did not automatically produce better performance. The effect depended on both the augmentation branch and the classifier.

LR benefited steadily from added data, improving from a no-augmentation Macro F1 of 0.494 to 0.576 with BackTranslation at 5x. For SVM, gains were smaller because the baseline was already strong, but the pattern remained meaningful: BackTranslation rose to 0.680 at 5x, SynonymAug peaked at 3x (0.679), and ContextualAug peaked earlier at 2x (0.676). RF improved more modestly and responded best to moderate or sentence-level augmentation, reaching 0.599 with BackTranslation at 5x.

BackTranslation likely worked best because it introduced sentence-level variation while preserving audit meaning more reliably than local word replacement. ContextualAug did not outperform it, probably because short audit findings concentrate class cues in a few domain terms, so even plausible local substitutions can soften or shift the clause signal. The remaining difficulty appears to reflect a combination of class imbalance, the very small Context of the Organization class, and genuine overlap between adjacent ISO categories rather than data volume alone.

The stronger performance of BackTranslation is also consistent with the feature representation used in this study. Because the models relied on TF-IDF unigrams and bigrams, useful gains were more likely when augmentation introduced alternative sentence forms while still preserving clause-bearing terms and short phrases. ContextualAug could generate locally plausible substitutions, but plausibility alone did not guarantee that the resulting wording remained discriminative in the sparse feature space. This helps explain why ContextualAug improved performance but peaked earlier, whereas BackTranslation continued to add value as the training set grew.

The size-response patterns also argue against a simple “more is always better” effect. Under SVM, SynonymAug reached its highest mean Macro F1 at 3x and then declined slightly at 5x, while ContextualAug peaked at 2x and softened thereafter. Only BackTranslation continued to improve up to 5x, and even there the margin over 3x remained modest. This non-monotonic behavior suggests that the reported gains were not produced by synthetic volume alone and reduces, although it does not eliminate, concern that the best 5x result reflects overfitting to repetitive augmented variants.

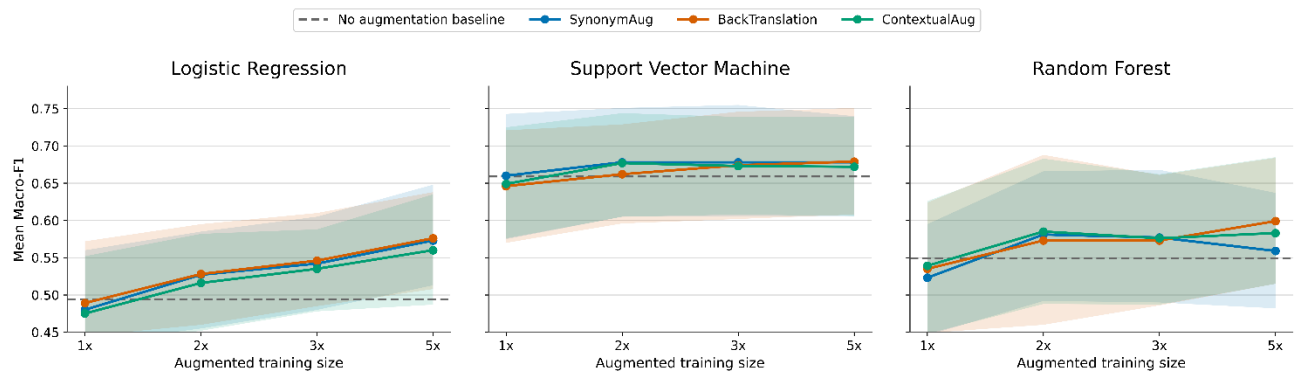


Figure 4. Synthetic data

4.4 Best Model Diagnostics

The best overall configuration was BackTranslation + SVM at 5x. Figure 5 shows its normalized confusion matrix.

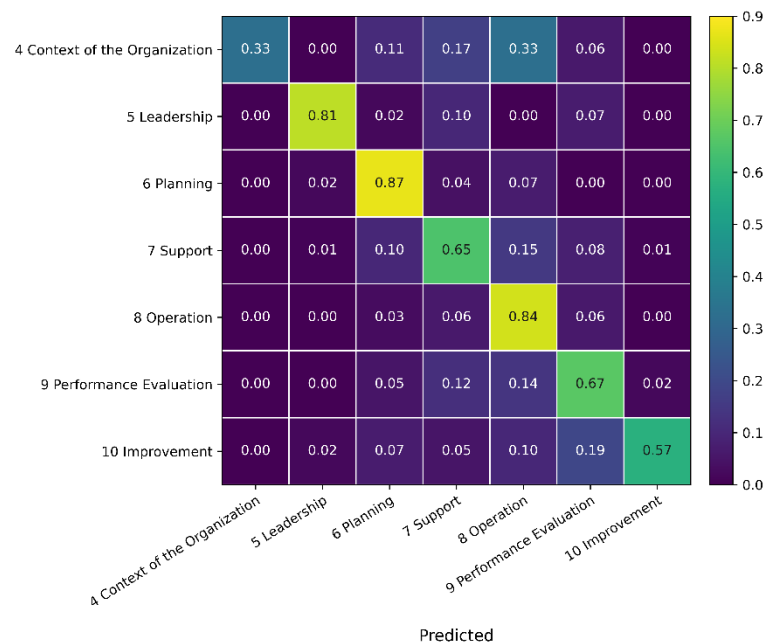


Figure 5. Normalized confusion matrix

Planning, Operation, and Leadership were the clearest classes, with diagonal values of 0.87, 0.84, and 0.81 respectively. Context of the Organization was the weakest class at 0.33, which is unsurprising given that it contained only six original findings.

The remaining errors were not random. Support-like findings often used phrases such as “training needs,” “assigning and allocating trainers,” and “capturing attendance,” whereas Performance Evaluation-like findings contained cues such as “formal objectives and KPIs,” “feedback analysis,” and “periodic reviews.” Operation-like findings included terms such as “Project Quality Plan” and “implemented in the execution.”

Some findings mixed these signals in the same record. For example, “capturing training feedback and analysis” blends Support with Performance Evaluation, while “SMART quality objectives ... implemented in the execution” blends Performance Evaluation with Operation. The augmentation examples show the same sensitivity: SynonymAug changed “due for review” to “pending assessment” and “Project Quality Plan” to “Project Quality framework,” shifts that can blur clause boundaries. This pattern suggests that the residual error comes from real clause overlap as well as model limitations.

At the same time, not all residual error should be attributed to augmentation quality. Some audit findings are genuinely multi-faceted: they can describe missing training evidence, weak monitoring, and incomplete implementation within a single statement. In such cases, confusion between Support, Performance Evaluation, and Operation may reflect the operational labeling scheme as much as the classifier. The results therefore point to three interacting sources of difficulty: class imbalance, real semantic proximity between adjacent clauses, and some unavoidable boundary softness in the labels. Future gains may therefore require not only better augmentation, but also richer annotation rules or multi-label treatment for borderline findings.

This is more than a modeling issue; it reflects the structure of the audit task itself. In practice, the operational distinction between these clauses is partly procedural and partly documentary. A single finding can mention training records, KPI monitoring, project execution, and corrective action in the same short narrative. When such signals co-occur, a single-label formulation forces the model to select a dominant clause even though the text carries evidence relevant to more than one category. The concentrated confusion seen in the matrix is therefore consistent with the audit process and not simply with classifier weakness. It suggests that future improvement may depend as much on refined annotation rules and reviewer guidance as on additional synthetic data.

Although the final corpus of 191 findings is small for a seven-class NLP classification task, the literature does not support a universal minimum threshold for such studies. Instead, small-sample classification is treated as a setting in which useful comparative evaluation remains possible, although performance estimates carry greater uncertainty and should be interpreted cautiously (Beleites et al., 2013; Varoquaux, 2018).

External validation on independent datasets remains limited in the present study. In prediction-model research, broader transportability requires evaluation beyond the development sample, while expert-reviewed annotation improves the credibility of the reference standard without substituting for independent testing (Steyerberg and Harrell, 2016; Artstein and Poesio, 2008). In this study, a randomly selected subset of clause assignments was reviewed by certified internal lead auditors together with an experienced internal auditor, which strengthened internal label validity, but further testing on later audit cycles, additional business units, or independent audit corpora is still required before broader generalization can be claimed.

Potential overfitting risk also remains a valid caution because augmentation reached 5x on a 191-record corpus. However, augmentation research in low-resource NLP consistently shows that gains are task-dependent rather than uniformly monotonic, and therefore larger synthetic sets should be treated as empirical test conditions rather than as assumed improvements (Chen et al., 2023; Pellicer et al., 2023). In the present study, the best 5x result is therefore interpreted as a dataset-specific optimum within this validation design, not as evidence that 5x augmentation is generally preferable.

4.5 Proposed Practical Improvement

The selected model can be moved from offline evaluation into a simple management-facing dashboard. Rather than replacing auditor judgment, it can support triage by assigning a likely ISO clause to each new finding, surfacing clause hotspots, and highlighting recurring language patterns.

As illustrated in Figure 6, the dashboard can combine clause hotspots, clause-level model performance, and frequent clause phrases in one view. Clear cases can be routed quickly to the relevant owner, while ambiguous or low-confidence cases can be flagged for manual review.

In practice, such a dashboard would be most useful as a decision-support layer around the audit process rather than as a stand-alone classifier. Predictions could be paired with confidence thresholds, allowing straightforward cases to be routed automatically while uncertain cases are reviewed by quality staff. Over time, confirmed corrections from users could also be stored as new labeled examples, gradually strengthening the training corpus. This would create a manageable feedback loop in which the model assists prioritization while the audit team continues to refine the data on which the model depends.

Implementation would also require basic governance around model use. The dashboard should expose low-confidence cases, retain the original finding text beside the predicted clause, and allow reviewers to correct outputs when needed. Those corrections are operationally valuable because they create a continuing stream of verified examples for future retraining. The same interface could also help management distinguish between stable clause hotspots and temporary reporting effects by linking predicted categories to repeated language cues over time. In that sense, deployment is not the end of the modeling process; it is part of the mechanism through which the audit function can gradually improve both its labeled data and its analytic support tools. Given the absence of external validation on an independent audit corpus, the first deployment stage should be a monitored decision-support pilot rather than a fully automated production workflow (Figure 6).

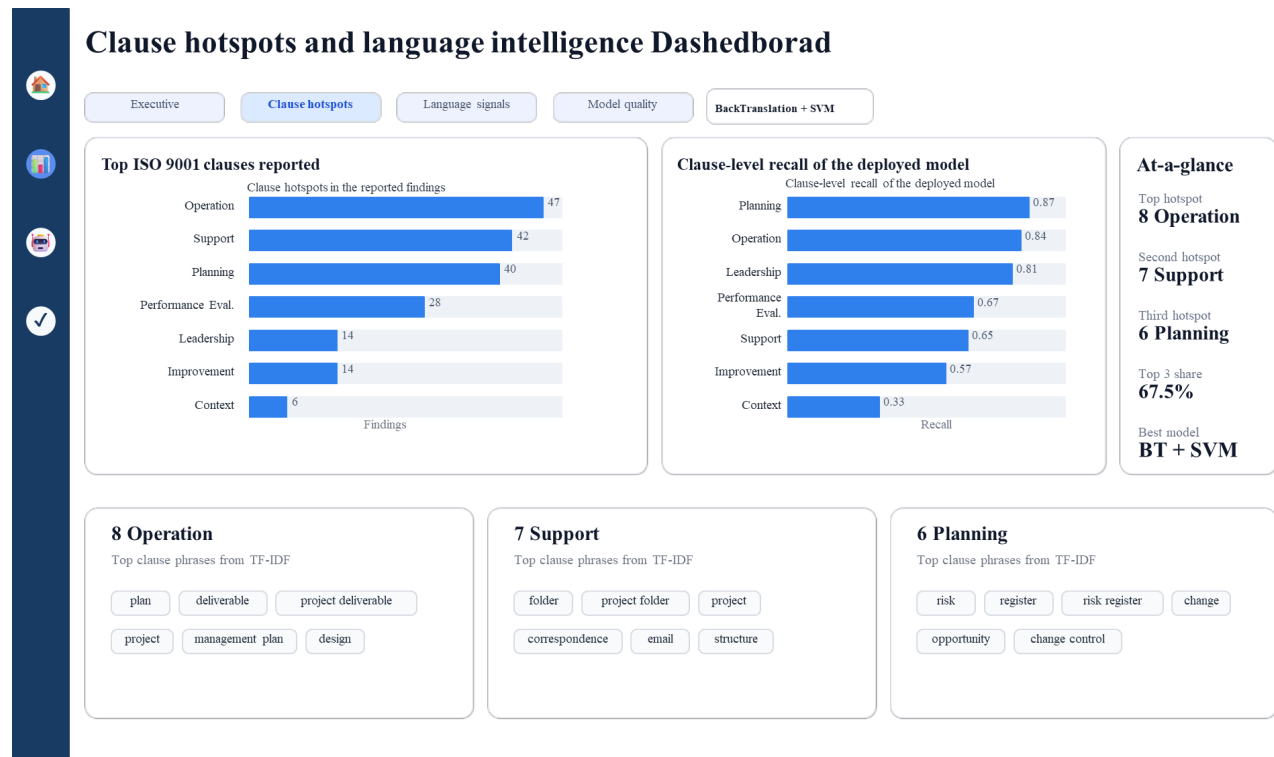


Figure 6. Proposed management dashboard for clause hotspots, clause-level model monitoring, and audit-language intelligence.

5. Conclusion

This study examined whether text augmentation improves clause-level classification of internal quality-audit findings under small and imbalanced labeled data. Controlled augmentation helped, but the gains were modest and depended on both the augmentation branch and the classifier.

Within this experiment, the strongest configuration was BackTranslation + SVM at 5x. The result suggests that sentence-level paraphrasing preserved audit meaning more effectively than local substitution, while the remaining errors reflected class imbalance and overlap between adjacent clauses, especially Support, Operation, and Performance Evaluation.

The study contributes a practical comparison framework for small audit-text datasets and a deployment path for dashboard-based triage. These results identify the most effective configuration within this small and imbalanced dataset, but broader generalization should be established through external validation on larger and more diverse audit corpora. Future work should also document augmentation hyperparameters more explicitly and test stronger drift checks. The model should therefore be viewed as strongly internally validated but only provisionally generalizable until it is tested on temporally later or genuinely independent audit datasets.

Although BackTranslation + SVM at 5x was the strongest configuration in this experiment, the broader pattern showed that higher augmentation ratios were not uniformly beneficial across branches. The best 5x result should therefore be read as a dataset-specific optimum within the present validation design rather than as a universal recommendation for small audit-text classification.

References

- Artstein, R. and Poesio, M., Inter-coder agreement for computational linguistics, *Computational Linguistics*, vol. 34, no. 4, pp. 555-596, 2008.
- Ansari, G., Garg, M. and Saxena, C., Data augmentation for mental health classification on social media, *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pp. 152-161, Silchar, India, 2021.
- Bayer, M., Kaufhold, M. A., Buchhold, B., Keller, M., Dallmeyer, J. and Reuter, C., Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers, *International Journal of Machine Learning and Cybernetics*, vol. 14, pp. 135-150, 2023.
- Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C. and Popp, J., Sample size planning for classification models, *Analytica Chimica Acta*, vol. 760, pp. 25-33, 2013.
- Chen, J., Tam, D., Raffel, C., Bansal, M. and Yang, D., An empirical survey of data augmentation for limited data learning in NLP, *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 191-211, 2023.
- Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T. and Hovy, E., A survey of data augmentation approaches for NLP, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 968-988, Online, 2021.
- International Organization for Standardization, *ISO 9001:2015, Quality management systems — Requirements*, 5th Edition, ISO, Geneva, Switzerland, 2015.
- Kapusta, J., Držik, D., Šteflovíč, K. and Nagy, K. S., Text data augmentation techniques for word embeddings in fake news classification, *IEEE Access*, vol. 12, pp. 31538-31550, 2024.
- Kobayashi, S., Contextual augmentation: data augmentation by words with paradigmatic relations, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 452-457, New Orleans, Louisiana, USA, 2018.
- Kumar, V., Choudhary, A. and Cho, E., Data augmentation using pre-trained transformer models, *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pp. 18-26, Online, December 7, 2020.
- Moreo, A., Esuli, A. and Sebastiani, F., Distributional random oversampling for imbalanced text classification, *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 805-808, Pisa, Italy, July 17-21, 2016.
- Pellicer, L. F. A. O., Ferreira, T. M. and Costa, A. H. R., Data augmentation techniques in natural language processing, *Applied Soft Computing*, vol. 132, article 109803, 2023.
- Shleifer, S., Low resource text classification with ULMFit and backtranslation, *arXiv preprint arXiv:1903.09244*, 2019.
- Steyerberg, E. W. and Harrell, F. E., Jr., Prediction models need appropriate internal, internal-external, and external validation, *Journal of Clinical Epidemiology*, vol. 69, pp. 245-247, 2016.

- Varoquaux, G., Cross-validation failure: Small sample sizes lead to large error bars, *NeuroImage*, vol. 180, pp. 68-77, 2018.
- Vaishampayan, S., Ramrakhiyani, N., Pawar, S., Pawde, A., Apte, M. and Palshikar, G. K., Audit report coverage assessment using sentence classification, *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pp. 31-41, Bali, Indonesia, 2023.
- Wang, D., Yin, K. and Wang, H., Intelligent classification of construction quality problems based on unbalanced short text data mining, *Ain Shams Engineering Journal*, vol. 15, no. 10, article 102983, 2024.
- Wang, Y., Zhang, Z., Wang, Z., Wang, C. and Wu, C., Interpretable machine learning-based text classification method for construction quality defect reports, *Journal of Building Engineering*, vol. 89, article 109330, 2024.
- Wei, J. and Zou, K., EDA: Easy data augmentation techniques for boosting performance on text classification tasks, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6382-6388, Hong Kong, China, 2019.

Biography

Husam Alghamdi is a Ph.D. candidate in Industrial Engineering at King Saud University, Riyadh, Saudi Arabia. He has more than 12 years of experience in quality management, QHSE, and operational excellence across manufacturing, oil and gas, and facility management. He holds a bachelor's degree in Mechanical Engineering from King Fahd University of Petroleum and Minerals and a master's degree from Alfaisal University. His research interests include quality management systems, ISO standards, governance, compliance, and data-driven process improvement. His current work focuses on applying analytics and machine learning to operational quality-management problems and audit-support systems.