Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 4 (2026) 898 - 908 e-ISSN: 2580-0760

Social Feature Integration for Entertainment Hoax Detection: Machine Learning and DistilBERT-Fusion

Dodo Zaenal Abidin[1], Agus Siswanto[2], Chindra Saputra[3], Bhetantio[4]
[1,4] Magister of Information System, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia
[2,3]Informatics Engineering, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia

*dodozaenalabidin@gmail.com

Abstract

Detecting entertainment hoaxes remains a pressing challenge in the social media era, where short texts, provocative headlines, and rapid dissemination complicate verification. This study compares the effectiveness of classical machine learning models and the transformer DistilBERT, both with and without the integration of social features. The modified GossipCop dataset, consisting of 22,140 entries, includes news titles as the primary text representation and social features such as tweet count, tweet density, and viral indicators. Text was represented using TF-IDF for classical models and DistilBERT tokenization for the transformer, with performance evaluated under stratified 10-fold cross-validation. Results show that incorporating social features consistently improves classical models, with XGBoost + Social achieving the best performance (PR-AUC 0.84; F1-score 0.76), surpassing DistilBERT-Fusion (PR-AUC 0.79). McNemar's test confirmed significant differences in error distributions, strengthening the reliability of these findings. To reduce reliance on a single empirical dataset, the social feature integration pipeline was additionally validated on the PolitiFact benchmark from a different domain, where incorporating minimal social signals yielded measurable and statistically significant gains over text-only models (Wilcoxon p = 0.0137), confirming the robustness of the social feature effect beyond the entertainment domain. Overall, the results highlight that for short-text entertainment news under the conditions examined in this study, boosting models enriched with social signals can outperform transformer-based approaches. While DistilBERT-Fusion provided competitive results, its improvement over text-only DistilBERT remained limited due to the short-text nature of the dataset, indicating opportunities for richer fusion strategies in future research.

Keywords: DistilBERT; hoax detection; GossipCop; machine learning; social features

How to Cite: D. Zaenal Abidin, A. Siswanto, C. Saputra, and B. Bhetantio, "Social Feature Integration for Entertainment Hoax Detection: Machine Learning and DistilBERT-Fusion", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 4, pp. 898 - 908, Jul. 2026.
Permalink/DOI: https://doi.org/10.29207/resti.v10i4.7208

Received: September 27, 2025
Accepted: June 30, 2026
Available Online: July 23, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The development of information technology and the increasing use of social media have fundamentally transformed the way society accesses, consumes, and disseminates information. Social media enables news to spread rapidly and broadens public participation in online discourse [1], [2]. At the same time, however, it also facilitates the dissemination of false information or hoaxes, which can lead to serious consequences such as declining public trust in the media, heightened potential for conflict, and the formation of opinions that deviate from facts [3-5].

One of the categories most vulnerable to hoaxes is entertainment news. Such news is often presented with short, provocative, or sensational headlines, making it difficult to distinguish between valid and misleading information [6], [7]. The limited textual representation constrains text-based classification models from fully capturing context, while the lightweight and viral nature of entertainment news further increases the risk of misinformation. This issue is compounded by class imbalance, where valid news significantly outnumbers false news, potentially biasing the performance of machine learning models [8], [9].

Previous research has attempted to address these challenges through both classical machine learning and advanced deep learning approaches. Shu et al. [7] introduced FakeNewsNet, which contains news content, social context, and temporal information from PolitiFact and GossipCop, and has since become a key reference for hoax detection research across domains. Benchmark studies by Khan et al. [8] demonstrated that transformer models such as BERT and RoBERTa

consistently outperform traditional methods on large and diverse datasets [10]. Other studies have focused on specific domains, such as Alghamdi et al. [11], who demonstrated the effectiveness of CT-BERT on health news, and Farooq et al. [12], who highlighted the potential of classical ensemble methods for Urdu-language news. These findings underline that the effectiveness of detection methods is highly dependent on the data context and research domain [13].

Although recent literature emphasizes the dominance of transformer models, the integration of social engagement features has received less attention. Social features, such as tweet count, distribution density, and viral indicators, provide additional insights into patterns of public interaction that cannot be captured by text alone. Prior studies Shu et al. [7] also showed that social and emotional dimensions significantly enhance hoax detection accuracy. Nevertheless, most research remains limited to textual analysis, and the potential of incorporating social features has not been fully explored, particularly in the entertainment domain [14], [15].

Building on this gap, the present study offers a comprehensive comparison between classical machine learning and the transformer model DistilBERT enriched with social features, using the GossipCop entertainment dataset. The main contributions of this study are threefold: (1) to evaluate the extent to which social feature integration improves the performance of classical machine learning models, (2) to compare the effectiveness of XGBoost and DistilBERT-Fusion on short-text news, and (3) to examine the influence of data characteristics on hoax detection performance. In addition, the study employs statistical evaluations using Wilcoxon and McNemar tests to ensure the validity of comparisons [16], [17]. This combination complements metric-based cross-validation by providing statistical evidence at both fold and instance levels. Wilcoxon evaluates whether performance gains are consistently observed across paired folds, while McNemar assesses whether models differ in their instance-level error patterns. Together, they make the comparison more robust to fold variability and threshold effects, which is particularly important in short-text, imbalanced hoax detection. Thus, this research provides theoretical contributions by expanding the understanding of cross-paradigm hoax detection methods and practical contributions by supporting the development of more adaptive and efficient entertainment fact-checking systems.

2. Methods

This section presents the end-to-end methodology employed in the study, summarized in Figure 1. The workflow begins with dataset loading and cleaning, followed by stratified 10-fold cross-validation to preserve balanced class distributions and ensure representative evaluation [18]. Within each fold, social feature engineering is applied independently to prevent data leakage, and text is represented using two approaches: TF-IDF for classical machine learning models and DistilBERT tokenization for the transformer. Each model is trained and validated on its corresponding fold, with predictions aggregated into out-of-fold (OOF) results for comprehensive evaluation using metrics suitable for imbalanced data [19].

Finally, performance analysis is complemented with statistical tests, including the Wilcoxon signed-rank and McNemar's tests, to ensure the validity, reliability, and significance of performance differences between models [17]. This framework enables a fair comparison across paradigms while systematically assessing the contribution of social features to effective hoax detection in entertainment news.

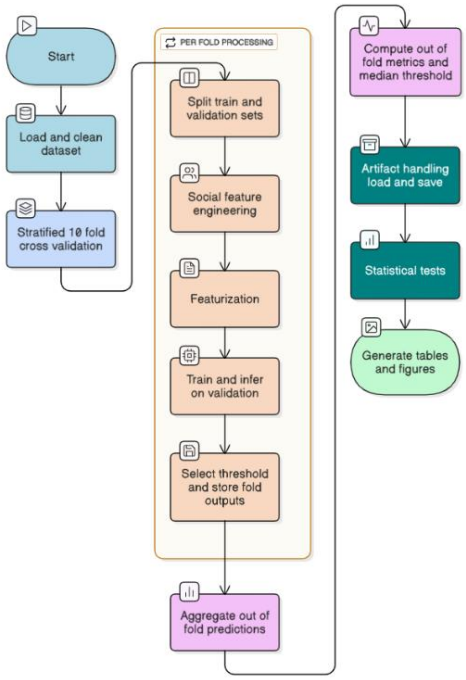


Figure 1. Workflow pipeline for GossipCop hoax detection

As shown in Figure 1, each stage of the workflow complements the others to ensure a valid and unbiased evaluation. Data cleaning provides consistent inputs, while stratified 10-fold cross-validation maintains balanced partitions. Social feature engineering is performed per fold to enrich textual information without leakage. Text is represented through TF-IDF and DistilBERT tokenization, enabling paradigm comparison. Predictions are aggregated into OOF results, with metrics and statistical tests applied and results presented in tables and figures to support analysis.

2.1 Overview of the Proposed Workflow

Figure 1 presents the overall workflow, which integrates key stages from data preparation to evaluation. Dataset cleaning ensures input quality, while stratified 10-fold cross-validation maintains balanced class distributions [20]. Social feature engineering is performed per fold to enrich information and avoid data leakage. Text is represented through two

approaches—TF-IDF for classical models and DistilBERT tokenization for transformer models—allowing cross-paradigm comparison [21], [22].

Model outputs are aggregated into out-of-fold predictions and evaluated using metrics suitable for imbalanced data, with statistical tests applied to confirm the significance of performance differences. This framework underpins the detailed explanations in the following subsections and highlights the study's objective of assessing the role of social features in enhancing hoax detection effectiveness.

2.2 Dataset Description and Preprocessing

The dataset used in this study is derived from the GossipCop corpus within the FakeNewsNet repository [7], comprising 22,140 entries in the entertainment news domain. Each entry includes a news title as the primary textual representation and several social features such as title_length, tweet_count, tweet_density, and binary_viral. The class label is binary (0 = valid, 1 = hoax), with an imbalanced distribution of 76% valid and 24% hoax news. This imbalance requires evaluation metrics sensitive to minority classes, such as PR-AUC and F1-score.

Preprocessing was conducted to ensure data quality and mitigate bias. Missing values and duplicate entries were removed, and numerical features were transformed for stability. For example, tweet_count was log-normalized to reduce skewness, while tweet_density was derived from the ratio of tweet volume to title length, log-transformed, winsorized at the 99.5th percentile, and normalized for scale. The binary_viral feature was created by applying a 90th percentile threshold of tweet counts in each fold. Importantly, all feature engineering was performed independently within the training partition of each stratified 10-fold cross-validation to prevent data leakage and ensure robust evaluation [23]. A concise summary of these preprocessing steps is provided in Table 1.

Table 1. Summary of Dataset Preprocessing

Step	Transformation & Purpose
Data cleaning	Remove missing values and duplicates → Ensure data quality
Text representation (n-gram features)	TF-IDF with unigram–bigram features → Capture word pairs including implicit negation patterns (e.g., "not good", "not bad") via bigram tokenization
tweet_count transformation	Log normalization → Reduce distribution skewness
tweet_density calculation	Ratio of tweets to title length → Normalize interaction signal
binary_viral creation	90th percentile threshold → Capture viral patterns
Per-fold feature engineering	Applied only on training set → Prevent data leakage

2.3 Cross-Dataset Validation on PolitiFact

To address potential dataset-specific bias and respond to concerns about relying solely on a single empirical dataset, this study additionally validates the proposed evaluation pipeline on the PolitiFact corpus within the FakeNewsNet repository [7]. The processed PolitiFact dataset contains 983 entries in the political news domain, with a binary label scheme consistent with the main experiments (0 = valid, 1 = hoax). This additional dataset supports cross-dataset assessment under a different topical distribution and a smaller-scale setting while preserving the same modeling and evaluation design.

For consistency with the GossipCop setting, each PolitiFact entry uses the news title as the primary textual representation. Minimal social information is available through tweet identifiers (tweet_ids), from which tweet_count is derived as the number of associated tweets. Following the preprocessing rationale in Section 2.2, tweet_count is log-normalized (log1p) to improve numerical stability. The title_length feature is computed to support normalization of interaction signals. Using these variables, tweet_density is constructed analogously to the main dataset (i.e., tweet volume normalized by title length and log-transformed), with fold-wise capping/winsorization applied to control extreme values. The binary_viral indicator is defined using a 90th percentile threshold of tweet_count estimated from the training split in each fold. Importantly, all fold-dependent feature engineering (e.g., density caps and viral thresholds) is performed strictly within the training partition of each stratified 10-fold cross-validation to prevent data leakage and ensure fair evaluation [23].

All experiments on PolitiFact follow the same stratified 10-fold cross-validation protocol and evaluation metrics used for GossipCop, including PR-AUC, F1-score (hoax class), MCC, and ROC-AUC. Overall, this cross-dataset validation is intended to verify whether the benefits of integrating social signals into short-text hoax detection remain stable beyond the entertainment domain, without altering the core methodological framework.

2.4 Cross-Validation Strategy

A stratified 10-fold cross-validation scheme was adopted to ensure robust evaluation under imbalanced class conditions. Stratification preserves the proportion of class 0 (valid news) and class 1 (hoax) in each fold, allowing model performance to be assessed consistently against the original distribution [23], [24].

All social feature engineering, such as the construction of binary_viral from the 90th percentile of tweet counts, was performed exclusively on the training partition of each fold. This procedure prevents data leakage and guarantees that validation data remain unseen during feature construction. Predictions from each fold were aggregated into out-of-fold (OOF) results, which served as the basis for metric computation and comparative analysis [25]. This strategy improves the reliability of evaluation, minimizes bias from class imbalance, and provides a fair foundation for comparing models across different paradigms.

2.5 Social Feature Engineering

In addition to text-based representation, this study incorporates social features to capture user interaction dynamics with news articles [26], [27]. Such features are crucial because dissemination patterns on social media often provide complementary signals for distinguishing valid news from hoaxes. Two key features are employed: binary_viral and tweet_density.

The binary_viral feature is derived from the distribution of tweet counts in the training data of each fold. Entries with tweet counts at or above the 90th percentile are labeled as viral, while the rest are categorized as non-viral. The tweet_density feature represents dissemination intensity relative to the length of the news title. It is calculated by dividing the number of tweets by the title length, applying a logarithmic transformation to reduce skewness, and normalizing the result to align with the input scale of machine learning models.

All social feature engineering is performed exclusively on the training set within the stratified 10-fold cross-validation scheme to prevent data leakage. The integration of binary_viral and tweet_density enriches the dataset with meaningful social signals, enabling models to better capture dissemination patterns that are especially relevant in the entertainment news domain.

2.6 Text Featurization

Following preprocessing and social feature construction, the next step is building text representations as the primary model input [28]. The quality of text representation is critical, as it directly influences the model's ability to distinguish valid from fake news. Two approaches are employed: Term Frequency–Inverse Document Frequency (TF-IDF) for classical models and DistilBERT tokenization for transformer-based models [29], [30].

For classical models, news titles are vectorized using TF-IDF with an n-gram range of 1–2 and a minimum word frequency (min_df = 2). The 1–2 n-gram range is optimal for short entertainment headlines because it captures both individual keywords and meaningful word pairs while avoiding sparsity from longer n-grams. Setting min_df = 2 reduces noise from rare terms that contribute little to classification.

For transformer-based models, text is processed with the DistilBERT tokenizer, which segments headlines into subword tokens with a maximum length of 96. This parameter aligns with the dataset characteristics, as nearly all titles fall within this limit, ensuring efficient computation while preserving contextual information. The resulting tokenized outputs are converted into contextual embeddings that capture semantic relationships more effectively than frequency-based methods.

By combining these two approaches, the study enables a direct comparison between traditional frequency-based and contextual representations. TF-IDF offers efficiency and interpretability, whereas DistilBERT provides richer semantic modeling that may improve detection of subtle hoax patterns [31], [32]. This dual setup ensures a comprehensive evaluation of text representation strategies in entertainment hoax detection.

2.7 Models Evaluated

This study evaluates two groups of algorithms—classical machine learning and transformer-based models—to assess their effectiveness in entertainment hoax detection.

For the classical group, three baseline models were tested with TF-IDF features: Logistic Regression (LR), Linear Support Vector Machine (SVM), and Complement Naive Bayes (CNB), chosen for their simplicity, interpretability, and common use in short-text classification [33]. In addition, two gradient boosting ensembles, Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM), were included due to their strong performance on high-dimensional and imbalanced data, as shown in prior hoax detection studies [34].

For the transformer group, DistilBERT was selected as a lightweight yet context-rich variant of BERT [35]. Two configurations were implemented: DistilBERT-Text, using only news titles, and DistilBERT-Fusion, which integrates textual embeddings with four social features (log_tweet_count, tweet_density, binary_viral, and title_length). The latter highlights the study's novelty by combining semantic representations with social engagement signals.

The DistilBERT-Fusion architecture operates as follows. The backbone model (distilbert-base-uncased) encodes each news title into a 768-dimensional CLS token representation. In parallel, four social features (log_tweet_count, tweet_density, binary_viral, and title_length) are processed through a dedicated MLP branch consisting of a Linear(4→64) layer, ReLU activation, and Dropout(0.1), producing a 64-dimensional social embedding. Prior to the MLP branch, social features were standardized using StandardScaler fitted exclusively on the training partition of each fold to prevent data leakage. The two representations are then concatenated to form a 832-dimensional fused vector, which is passed through a Dropout(0.1) layer and a final Linear(832→2) classifier head. All parameters, including the DistilBERT backbone, were fully fine-tuned end-to-end using AdamW.

This evaluation framework enables a comprehensive comparison, from efficient baselines to advanced transformers, while emphasizing the contribution of social features in enhancing detection performance.

2.8 Training and Validation Protocol

Training and evaluation followed a stratified 10-fold cross-validation scheme, ensuring that results reflect model generalization across the corpus [36]. The

decision threshold was optimized through a Precision–Recall sweep (PR–F1 sweep) on validation data, selecting the point that balanced precision and recall to improve hoax detection under imbalanced conditions.

For DistilBERT-based models, training was conducted with a learning rate of 2×10^{-5}, batch size of 32, and 3 epochs per fold. A linear warmup schedule covering 10% of total training steps was applied, followed by linear decay. Gradient norms were clipped at 1.0 to stabilize training. These settings were applied consistently across all 10 folds for both DistilBERT-Text and DistilBERT-Fusion configurations.

Predictions from all folds were aggregated into out-of-fold (OOF) results, providing an unbiased basis for global evaluation since each instance was predicted by a model unseen during training. Evaluation employed PR-AUC as the primary metric, complemented by F1-score for the hoax class, Matthews Correlation Coefficient (MCC), and ROC-AUC [37].

This protocol ensured fair, leakage-free evaluation while enabling advanced statistical validation. Wilcoxon signed-rank and McNemar tests were applied to OOF predictions to confirm the significance of observed performance differences.

2.9 Evaluation Metrics

Model performance was primarily assessed using the Precision–Recall Area Under the Curve (PR-AUC), which is more informative than ROC-AUC under imbalanced conditions [21]. PR-AUC highlights the trade-off between precision and recall, making it particularly suitable for hoax detection where errors on the minority class carry greater consequences [38].

To provide complementary perspectives, three additional metrics were reported. The F1-score for the hoax class measures the balance between precision and recall at the minority level, while the Matthews Correlation Coefficient (MCC) incorporates all confusion matrix elements to offer a robust correlation measure even with skewed distributions. ROC-AUC was also included for consistency with previous studies and to illustrate overall class separability.

All metrics were calculated from out-of-fold (OOF) predictions produced by the cross-validation protocol, ensuring unbiased and reliable evaluation. Together, these metrics provide both a primary focus on minority detection effectiveness and a broader view of overall model performance.

2.10 Statistical Analysis

To ensure that performance differences were not driven by random variation, this study employed two non-parametric tests: the Wilcoxon signed-rank test and McNemar's test [39]. The Wilcoxon test assessed the impact of social features by comparing text-only models with text+social models across folds, making it suitable for paired evaluations of incremental improvements. Meanwhile, the McNemar test analyzed error distributions between the best-performing classical model (XGBoost + Social) and the transformer-based DistilBERT-Fusion, focusing on instances where their predictions diverged.

Together, these tests provide complementary perspectives: Wilcoxon validates the significance of social feature integration at the fold level, while McNemar captures distributional differences in classification errors between top models [40], [41]. This dual approach establishes a robust statistical basis for the findings presented in Chapter 3.

3. Results and Discussions

This chapter presents the experimental findings and their analysis to assess the effectiveness of different approaches for entertainment hoax detection. The discussion is structured into several parts: first, an overview of dataset characteristics to contextualize the results; second, the performance of baseline classical machine learning models; third, the added value of social feature integration; and fourth, the outcomes of transformer-based DistilBERT models. A comparative analysis between the best-performing classical and transformer approaches is then provided, reinforced by statistical testing and visualization-based interpretation. The chapter closes by highlighting the theoretical and practical implications of these findings in addressing the research objectives and guiding the development of adaptive hoax detection systems.

3.1 Dataset Characteristics

The dataset is drawn from the GossipCop corpus in the entertainment domain, comprising 22,140 entries with binary labels (0 = valid, 1 = hoax) and an imbalanced distribution of approximately 76% valid versus 24% hoax news [23]. As detailed in Section 2.2, preprocessing ensured data quality and introduced social features, including tweet_count, tweet_density, and binary_viral, constructed per fold to avoid leakage. These features complement textual content by capturing dissemination dynamics that are particularly relevant in social media contexts.

The integration of textual and social dimensions provides a richer basis for evaluating how classical machine learning and transformer-based models exploit complementary information sources. This dataset structure also underpins the interpretation of baseline results presented in the next subsection.

3.2 Baseline Machine Learning Performance

As an initial step, this study evaluated the performance of classical machine learning models using text-based representations with Term Frequency–Inverse Document Frequency (TF-IDF). Three baseline algorithms were employed: Logistic Regression (LR), Support Vector Machine (SVM), and Complement Naive Bayes (CNB), selected for their simplicity, efficiency, and widespread use in short-text

classification. The baseline results are presented in Table 2.

Table 2. Baseline performance of machine learning models using TF-IDF (text only).

Model	PR-AUC	F1 (Hoax)	MCC	ROC-AUC
Logistic Regression	0.7429	0.6824	0.5761	0.8843
Support Vector Machine	0.7462	0.6795	0.5796	0.8796
Complement Naive Bayes	0.7329	0.6712	0.5645	0.8726

All three models demonstrated moderate performance, with CNB producing the lowest scores (PR-AUC 0.7329), while LR and SVM performed slightly better with PR-AUC values of 0.7429 and 0.7462, respectively. MCC values ranging from 0.56 to 0.58 confirm that text-only baselines yield mid-level performance under imbalanced data conditions. This limitation is further reflected in the gap between relatively high ROC-AUC values (~0.87–0.88) and lower PR-AUC values (~0.73–0.75), indicating difficulty in identifying the minority class (hoaxes), which represents only 24% of the dataset.

Overall, these baseline findings emphasize the limitations of text-only models, which tend to bias toward the majority class. This justifies the subsequent evaluation of social feature integration to determine whether complementary signals can improve hoax detection, as discussed in Subsection 3.3.

3.3 Effect of Social Feature

This subsection examines the effect of integrating social features—tweet_count, tweet_density, and binary_viral—on classical machine learning models. These features were engineered per fold to prevent data leakage, ensuring fair evaluation. The comparison between text-only and text+social models is summarized in Table 3.

Table 3. Performance comparison of text-only and text+social models.

Model	PR-AUC	F1 (Hoax)	MCC	ROC-AUC
Logistic Regression	0.7429	0.6824	0.5761	0.8843
Logistic Regression + Social	0.7540	0.6970	0.5951	0.8984
Support Vector Machine	0.7462	0.6795	0.5796	0.8796
Support Vector Machine + Social	0.7655	0.6957	0.5946	0.8955
XGBoost	0.7179	0.6527	0.5365	0.8610
XGBoost + Social	0.8379	0.7583	0.6831	0.9220
LightGBM	0.6877	0.6313	0.5098	0.8415
LightGBM + Social	0.8289	0.7473	0.6681	0.9153

The integration of social features consistently improved all classical models across both linear and boosting paradigms. For linear models, PR-AUC increased from 0.7429 to 0.7540 for LR and from 0.7462 to 0.7655 for SVM, accompanied by higher F1-scores and MCC values. More notably, the boosting models demonstrated substantially larger gains: XGBoost improved from 0.7179 to 0.8379 (ΔPR-AUC = +0.120) and LightGBM from 0.6877 to 0.8289 (ΔPR-AUC = +0.141), indicating that gradient boosting methods benefit most from the addition of social signals, likely due to their capacity to exploit non-linear interactions between social features and sparse lexical cues. Statistical validation using the paired Wilcoxon signed-rank test confirmed the significance of these improvements for all four models (LR: $p = 0.0098$, $r = 0.790$; SVM: $p = 0.0020$, $r = 0.886$; XGBoost: $p = 0.0020$, $r = 0.886$; LightGBM: $p = 0.0020$, $r = 0.886$), with XGBoost and LightGBM showing consistent gains across all 10 folds. All effect sizes exceeded $r = 0.7$, indicating very large practical significance.

In practical terms, the results highlight that social engagement signals—such as viral intensity and interaction density—offer valuable complementary context beyond short headlines. For entertainment news, where texts are brief and often sensational, incorporating these signals enables models to better recognize hoax propagation patterns. Thus, social feature integration provides a necessary foundation before advancing to transformer-based approaches, discussed in the next subsection.

3.4 Transformer Models

Beyond classical baselines, transformer-based approaches were evaluated using DistilBERT, selected for its efficiency and ability to generate strong contextual embeddings while requiring significantly lower computational resources than the full BERT model. Two configurations were examined: DistilBERT-Text, which relies exclusively on news titles, and DistilBERT-Fusion, which enriches textual embeddings with complementary social features to capture interaction dynamics more effectively. This dual setup enables a more comprehensive evaluation of contextual versus hybrid representations, offering insights into the relative contributions of semantic and social information dimensions. The results are summarized in Table 4.

Table 4. Performance of transformer-based models using news titles from GossipCop.

Model	PR-AUC	F1 (Hoax)	MCC	ROC-AUC
DistilBERT-Text	0.7865	0.7045	0.6086	0.8975
DistilBERT-Fusion	0.7884	0.7073	0.6151	0.8987

Both configurations delivered balanced performance, with DistilBERT-Fusion achieving only marginal gains over DistilBERT-Text (PR-AUC 0.7884 vs. 0.7865). While social signals improved MCC slightly, the effect remained limited due to the brevity of news titles, which restricts the capacity of transformers to fully leverage interaction features.

Compared to TF-IDF baselines (PR-AUC ~0.74–0.75), DistilBERT clearly provided stronger handling of imbalanced data, benefiting from contextual representations. However, the findings also confirm that social feature contributions are less pronounced in

short-text scenarios, aligning with prior studies that noted the challenge of applying transformers to minimal textual input.

Overall, DistilBERT—both text-only and fusion—achieved competitive results, offering a robust reference point for the comparative analysis with classical models in the next subsection.

3.5 Comparative Analysis of Machine Learning and Transformer Models

This subsection compares the best-performing models from the two paradigms: gradient boosting with social feature integration and the transformer-based DistilBERT. The goal is to assess the relative strengths of traditional frequency-based representations enriched with social signals versus contextual embeddings. The results are presented in Table 5.

Table 5. Comparative performance of best-performing machine learning and transformer models on GossipCop dataset.

Model	PR-AUC	F1 (Hoax)	MCC	ROC-AUC
XGB + Social	0.8379	0.7583	0.6831	0.9220
LGBM + Social	0.8289	0.7473	0.6681	0.9153
DistilBERT-Fusion	0.7884	0.7073	0.6151	0.8987
DistilBERT-Text	0.7865	0.7045	0.6086	0.8975

The results show that XGB+Social delivered the strongest performance (PR-AUC 0.8379; F1-score 0.7583; MCC 0.6831), followed by LGBM+Social with similarly competitive outcomes. In contrast, DistilBERT-Fusion achieved lower scores (PR-AUC 0.7884; F1-score 0.7073), only slightly surpassing DistilBERT-Text. The margin of approximately 0.05 in PR-AUC between XGB+Social and DistilBERT-Fusion is considerable under imbalanced data conditions, underscoring the practical advantage of boosting with social features.

These differences were statistically validated using McNemar's test ($p < 0.001$), with error analysis showing that XGB+Social generated fewer misclassifications than DistilBERT-Fusion ($b = 1857$, $c = 1295$). This reinforces the reliability of the classical ensemble's advantage over DistilBERT-Fusion within this experimental setting.

Overall, the findings demonstrate that in short-text entertainment news, gradient boosting models enriched with social engagement signals outperform transformer-based approaches. This suggests that transformer superiority should not be assumed universal, at least under the short-text, social-feature-augmented conditions examined here, and underscores the importance of aligning model selection with dataset and domain characteristics. At the same time, the results open opportunities for hybrid or ensemble strategies that could combine the interpretability and efficiency of boosting with the contextual depth of transformer models.

3.6 Visualization and Error Analysis

To complement the numerical evaluation, visualization and error analysis are employed to provide deeper insights into model performance. Two main tools are used: Precision–Recall (PR) curves, which illustrate the trade-off between precision and recall, and confusion matrices, which show the distribution of correct and incorrect predictions. To further illustrate the trade-off between precision and recall, the PR curves are presented in Figure 2. A concise summary of critical errors in terms of false negatives (FN) and false positives (FP) is also reported in Table 6.

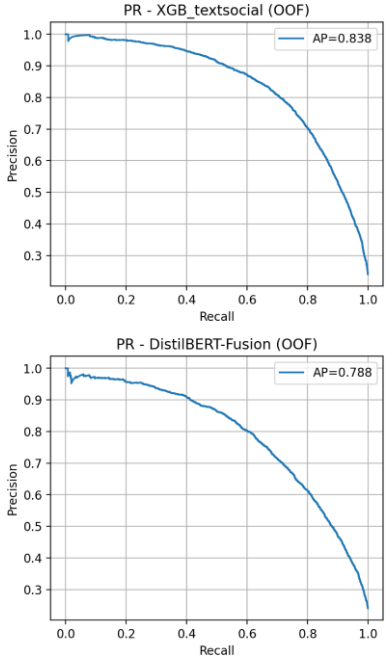


Figure 2. Precision–Recall curves for XGB+Social and DistilBERT-Fusion (OOF results)

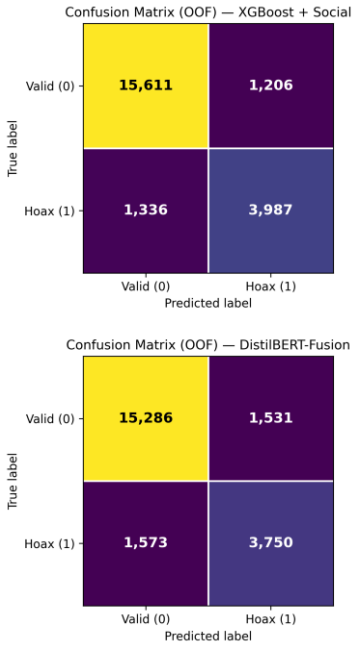


Figure 3. Confusion matrices for XGB+Social (top) and DistilBERT-Fusion (bottom)

As shown in Figure 2, XGB+Social achieves a wider PR area with AP = 0.838 compared to DistilBERT-Fusion with AP = 0.788. The curve of XGB+Social is more stable, reflecting a better balance between precision and recall, whereas DistilBERT-Fusion tends to drop recall more sharply when maintaining high precision. These findings align with the numerical metrics, confirming the advantage of boosting models supported by social features for short-text detection. To further examine how these differences manifest in specific error distributions, confusion matrices are presented in Figure 3.

The confusion matrices in Figure 3 complement the PR curves by visualizing error distributions. Table 6 summarizes these differences, offering a clear numerical comparison.

Table 6. Summary of false negatives (FN) and false positives (FP) for XGB+Social and DistilBERT-Fusion

Model	False Negative (FN)	False Positive (FP)
XGB+Social	1336	1206
DistilBERT-Fusion	1573	1531

Table 6 shows that XGB+Social produces fewer false negatives (1336 vs. 1573), which is crucial for hoax detection since missed hoaxes (FN) pose a greater risk in real applications. DistilBERT-Fusion slightly reduces false positives (1531 vs. 1206), which may help maintain user trust by avoiding excessive false alarms, yet this advantage is outweighed by its weaker recall of hoaxes.

Overall, the visualization and error analysis suggest that while transformers capture contextual semantics effectively, boosting models enriched with social features can be more reliable for headline-length entertainment hoax detection in this setting. This interpretation aligns with prior work emphasizing the value of social context in misinformation detection (e.g., Shu et al. [7]), where engagement-derived signals provide complementary evidence when textual cues are short and ambiguous. Importantly, this advantage is reflected in the reduction of critical errors—particularly false negatives—thereby lowering the likelihood of hoaxes going undetected and strengthening the practical utility of the proposed text+social configuration. These observations motivate the broader discussion in Section 3.8 on when and why social-context augmentation is most beneficial for short-text hoax detection.

3.7 Cross-Dataset Validation Results (PolitiFact)

We report cross-dataset validation results on PolitiFact to assess whether the performance trends observed on GossipCop remain consistent in a political-domain setting. The PolitiFact dataset contains 983 news titles (555 valid and 428 hoax instances) and was evaluated using the protocol described in Section 2.3. Table 7 summarizes the out-of-fold (OOF) performance under stratified 10-fold cross-validation for XGBoost in the text-only and text+social configurations.

Table 7. Performance Comparison on PolitiFact

Model	PR-AUC	F1 (Hoax)	MCC	ROC-AUC
XGB (Text)	0.7628	0.6979	0.4567	0.7897
XGB (Text+Social)	0.7991	0.7392	0.5217	0.8278

The results indicate that incorporating social features leads to consistent improvements across all reported metrics. PR-AUC increases from 0.7628 to 0.7991, while the F1-score for the hoax class improves from 0.6979 to 0.7392. MCC also rises from 0.4567 to 0.5217, indicating more reliable separation between valid and hoax instances under the same cross-validation setting. Based on the OOF confusion matrix, false negatives decrease from 123 (text-only) to 95 (text+social), suggesting improved detection of hoax items.

Although the PolitiFact dataset provides relatively sparse engagement information and a smaller sample size than GossipCop, incorporating these minimal social signals still yields measurable gains. This pattern is consistent with the improvements observed in Sections 3.2–3.6, where adding social features systematically strengthened performance in the entertainment domain. It should be noted, however, that this cross-dataset validation examines only the social feature effect using XGBoost; the transformer-based comparison is specific to the GossipCop setting and is not replicated here. Overall, the cross-dataset evidence suggests—within the scope of the present experimental setting—that social feature integration can enhance short-text hoax detection beyond a single dataset while maintaining an unchanged modeling and evaluation framework. Statistical validation using the paired Wilcoxon signed-rank test further confirmed that these gains are consistent across folds ($p = 0.0137$, effect size $r = 0.757$, 9/10 folds positive), indicating a large and reliable improvement from social feature integration on PolitiFact.

3.8 Discussions

The integration of social features had a clear and statistically supported impact on entertainment hoax detection in this study. Text-only models with TF-IDF achieved PR-AUC values of 0.69–0.75, whereas the addition of social features consistently improved all classical models ($p < 0.01$ for all four models). XGBoost with social features delivered the best performance (PR-AUC 0.8379; F1-score 0.7583), outperforming DistilBERT-Fusion (PR-AUC 0.7884), as further validated by McNemar's test. In addition to overall metric gains, the improvements are also meaningful from an error perspective, as social augmentation helps reduce missed hoaxes (false negatives) while maintaining a competitive precision–recall trade-off.

These results extend prior work emphasizing the role of social signals in misinformation detection. Shu et al. [7]

highlighted the importance of social context in FakeNewsNet, and our findings confirm its relevance in the entertainment domain. While Khan et al. [8] reported strong transformer performance on large and diverse datasets, the present results suggest that for short and sensational headlines, classical models enriched with social features can remain more effective. A plausible explanation is that headline-length inputs provide limited contextual evidence, which can constrain the benefit of contextual embeddings, whereas boosting methods can better leverage structured social signals (e.g., intensity and normalization of engagement) and their interactions with sparse lexical cues. This mechanism is particularly relevant under short-text conditions and class imbalance, where modest metric gains may correspond to substantial reductions in critical errors.

Importantly, the cross-dataset validation on PolitiFact further supports the robustness of this interpretation beyond a single empirical dataset. Despite differences in domain (political vs. entertainment) and relatively sparse engagement information, incorporating minimal social signals still yields measurable improvements under the same evaluation framework. This cross-dataset consistency helps clarify the data conditions in which the proposed text+social configuration is most advantageous: short-text news titles, limited textual context, and the availability of engagement-derived signals that provide complementary information to the headline content.

Practically, boosting models with social features offer efficient solutions for fact-checking systems in entertainment contexts with limited resources. Academically, this research underscores that transformer superiority should not be assumed universal, but may be contingent on input length, domain characteristics, and the strength of auxiliary signals, at least within the experimental settings tested here. Limitations include the use of headlines as the sole textual input and the relatively simple scope of social features. Future studies should explore richer engagement signals and hybrid or explainable AI approaches to enhance interpretability and robustness across broader settings.

Taken together, these findings provide both theoretical and practical contributions by clarifying when social-context augmentation is most beneficial and by supporting robust model selection for short-text hoax detection.

4. Conclusions

This study investigated entertainment hoax detection in short-text news by emphasizing the integration of social features into classification models. The results show that text-only TF-IDF approaches are limited when headlines provide minimal contextual information and the class distribution is skewed toward valid news, which hinders minority (hoax) detection. Incorporating social signals—such as engagement intensity and normalized dissemination indicators (e.g., density- and virality-based features)—consistently improved performance across all classical models, with particularly pronounced gains in the boosting models. XGBoost with social features achieved the best outcome and outperformed DistilBERT-Fusion in this experimental setting, indicating that boosting-based methods remain highly competitive for headline-length inputs where structured social cues provide complementary evidence, though these findings are specific to the tested configuration and dataset characteristics.

The primary contribution of this research is empirical evidence that transformer superiority is not guaranteed under conditions similar to those examined here, but may be conditional on data characteristics, particularly in short-text, imbalanced settings such as the one examined here. The proposed text+social configuration is most beneficial under short-title regimes, limited textual context, and the availability of engagement-derived signals that help resolve ambiguity in sensational or clickbait-like headlines. To reduce reliance on a single empirical dataset, we validated the same pipeline on an additional benchmark dataset (PolitiFact) from a different domain and observed consistent gains when minimal social signals were incorporated, supporting robustness beyond a single entertainment dataset.

Limitations include reliance on headlines as the sole textual input and the relatively simple design of social features. Future research should incorporate richer textual content and engagement signals, explore advanced social interaction metrics, and examine hybrid or explainable AI approaches to improve transparency and interpretability. Overall, this work contributes a practical and methodologically consistent step toward efficient and domain-aware hoax detection in real-world social media environments.

Acknowledgements

The author declares no conflict of interest. This work was supported by a research grant from Hibah KemdiktiSaintek 2025 [Grant No. 0070/C3/AL.04/2025]. The authors also gratefully acknowledge the support provided by the academic community of Universitas Dinamika Bangsa during the course of this research.

References

[1] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, “A review on sentiment analysis from social media platforms,” *Expert Syst. Appl.*, vol. 223, p. 119862, Aug. 2023, doi: 10.1016/j.eswa.2023.119862.

[2] B. T.K., C. S. R. Annavarapu, and A. Bablani, “Machine learning algorithms for social media analysis: A survey,” *Comput. Sci. Rev.*, vol. 40, p. 100395, May 2021, https://doi.org/10.1016/j.cosrev.2021.100395.

[3] M. Shahbazi and D. Bunker, “Social media trust: Fighting misinformation in the time of crisis,” *Int. J. Inf. Manag.*, vol. 77, p. 102780, Aug. 2024, doi: 10.1016/j.ijinfomgt.2024.102780.

[4] S. Muhammed T and S. K. Mathew, "The disaster of misinformation: a review of research in social media," *Int. J. Data Sci. Anal.*, vol. 13, no. 4, pp. 271–285, May 2022, doi: 10.1007/s41060-022-00311-6.

[5] N. Cavus, M. Goksu, and B. Oktekin, "Real-time fake news detection in online social networks: FANDC Cloud-based system," *Sci. Rep.*, vol. 14, no. 1, p. 25954, Oct. 2024, doi: 10.1038/s41598-024-76102-9.

[6] I. Ali, M. N. B. Ayub, P. Shivakumara, and N. F. B. M. Noor, "Fake News Detection Techniques on Social Media: A Survey," *Wirel. Commun. Mob. Comput.*, vol. 2022, no. 1, p. 6072084, Jan. 2022, doi: 10.1155/2022/6072084.

[7] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media," Mar. 27, 2019, *arXiv*: arXiv:1809.01286. doi: 10.48550/arXiv.1809.01286.

[8] J. Y. Khan, Md. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," *Mach. Learn. Appl.*, vol. 4, p. 100032, Jun. 2021, doi: 10.1016/j.mlwa.2021.100032.

[9] S. V. Balshetwar, A. Rs, and D. J. R, "Fake news detection in social media based on sentiment analysis using classifier techniques," *Multimed. Tools Appl.*, vol. 82, no. 23, pp. 35781–35811, Sep. 2023, doi: 10.1007/s11042-023-14883-3.

[10] M. Nasser *et al.*, "A systematic review of multimodal fake news detection on social media using deep learning models," *Results Eng.*, vol. 26, p. 104752, Jun. 2025, doi: 10.1016/j.rineng.2025.104752.

[11] J. Alghamdi, Y. Lin, and S. Luo, "Towards COVID-19 fake news detection using transformer-based models," *Knowl.-Based Syst.*, vol. 274, p. 110642, Aug. 2023, doi: 10.1016/j.knosys.2023.110642.

[12] M. S. Farooq, A. Naseem, F. Rustam, and I. Ashraf, "Fake news detection in Urdu language using machine learning," *PeerJ Comput. Sci.*, vol. 9, p. e1353, May 2023, doi: 10.7717/peerj-cs.1353.

[13] S. Yakkundi, R. Patil, S. Sangani, R. H. Goudar, S. I. Goudar, and A. Qazi, "Exploring machine learning for fake news detection: techniques, tools, challenges, and future research directions," *Discov. Appl. Sci.*, vol. 7, no. 8, p. 873, Aug. 2025, doi: 10.1007/s42452-025-07548-3.

[14] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," Sep. 03, 2017, *arXiv*: arXiv:1708.01967. doi: 10.48550/arXiv.1708.01967.

[15] D. Mouratidis, A. Kanavos, and K. Kermanidis, "From Misinformation to Insight: Machine Learning Strategies for Fake News Detection," *Information*, vol. 16, no. 3, p. 189, Feb. 2025, doi: 10.3390/info16030189.

[16] B. Şener, K. Acici, and E. Sümer, "Categorization of Alzheimer's disease stages using deep learning approaches with McNemar's test," *PeerJ Comput. Sci.*, vol. 10, p. e1877, Feb. 2024, doi: 10.7717/peerj-cs.1877.

[17] S. Uddin and H. Lu, "Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data," *PLOS ONE*, vol. 19, no. 4, p. e0301541, Apr. 2024, doi: 10.1371/journal.pone.0301541.

[18] M. Adnan, A. A. S. Alarood, M. I. Uddin, and I. Ur Rehman, "Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models," *PeerJ Comput. Sci.*, vol. 8, p. e803, Feb. 2022, doi: 10.7717/peerj-cs.803.

[19] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4845–4901, Jul. 2024, doi: 10.1007/s10994-022-06268-8.

[20] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion," *Heliyon*, vol. 9, no. 10, p. e20382, Oct. 2023, doi: 10.1016/j.heliyon.2023.e20382.

[21] D. Z. Abidin, L. Afuan, A. N. Toscany, and Nurhadi, "A Comprehensive Benchmarking Pipeline for Transformer-Based Sentiment Analysis using Cross-Validated Metrics," *J. Tek. Inform. Jutif*, vol. 6, no. 4, pp. 1797–1810, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.4.4894.

[22] P. Guleria, J. Frnda, and P. N. Srinivasu, "NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis," *Array*, vol. 27, p. 100467, Sep. 2025, doi: 10.1016/j.array.2025.100467.

[23] A. Tariq, A. Mehmood, M. Elhadef, and M. U. G. Khan, "Adversarial Training for Fake News Classification," *IEEE Access*, vol. 10, pp. 82706–82715, 2022, doi: 10.1109/ACCESS.2022.3195030.

[24] A. Chadha and B. Kaushik, "A Hybrid Deep Learning Model Using Grid Search and Cross-Validation for Effective Classification and Prediction of Suicidal Ideation from Social Network Data," *New Gener. Comput.*, vol. 40, no. 4, pp. 889–914, Dec. 2022, doi: 10.1007/s00354-022-00191-1.

[25] I. Mohiuddin and A. Almogren, "Ensemble techniques for detecting profile cloning attacks in online social networks," *PeerJ Comput. Sci.*, vol. 11, p. e3182, Sep. 2025, doi: 10.7717/peerj-cs.3182.

[26] R. Scharlach and B. Hallinan, "The value affordances of social media engagement features," *J. Comput.-Mediat. Commun.*, vol. 28, no. 6, p. zmad040, Sep. 2023, doi: 10.1093/jcmc/zmad040.

[27] S. K. Uppada, K. Manasa, B. Vidhathri, R. Harini, and B. Sivaselvan, "Novel approaches to fake news and fake account detection in OSNs: user social engagement and visual content centric model," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, p. 52, Dec. 2022, doi: 10.1007/s13278-022-00878-9.

[28] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A Survey of Text Representation and Embedding Techniques in NLP," *IEEE Access*, vol. 11, pp. 36120–36146, 2023, doi: 10.1109/ACCESS.2023.3266377.

[29] L. R. Aini and Evi Yulianti, "Expertise Retrieval Using Adjusted TF-IDF and Keyword Mapping to ACM Classification Terms," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 9, no. 3, pp. 497–505, May 2025, doi: 10.29207/resti.v9i3.6397.

[30] A. Hartanto, Ema Utami, Arief Setyanto, and Kusrini, "Evaluating Transformer Models for Social Media Text-Based Personality Profiling," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 9, no. 1, pp. 20–30, Jan. 2025, doi: 10.29207/resti.v9i1.6157.

[31] L. Zhang, "Features extraction based on Naive Bayes algorithm and TF-IDF for news classification," *PLOS One*, vol. 20, no. 7, p. e0327347, Jul. 2025, doi: 10.1371/journal.pone.0327347.

[32] S. Kofi Akpatsa *et al.*, "Online News Sentiment Classification Using DistilBERT," *J. Quantum Comput.*, vol. 4, no. 1, pp. 1–11, 2022, doi: 10.32604/jqc.2022.026658.

[33] J. Alghamdi, S. Luo, and Y. Lin, "A comprehensive survey on machine learning approaches for fake news detection," *Multimed. Tools Appl.*, vol. 83, no. 17, pp. 51009–51067, Nov. 2023, doi: 10.1007/s11042-023-17470-8.

[34] L. Jixian, A. Gang, S. Zhihao, and S. Xiaoqiang, "Social Media Multimodal Information Analysis based on the BiLSTM-Attention-CNN-XGBoost Ensemble Neural Network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 12, 2022, doi: 10.14569/IJACSA.2022.0131215.

[35] M. Hussain *et al.*, "Optimised knowledge distillation for efficient social media emotion recognition using DistilBERT and ALBERT," *Sci. Rep.*, vol. 15, no. 1, p. 30104, Aug. 2025, doi: 10.1038/s41598-025-16001-9.

[36] F. E. Arévalo-Cordovilla and M. Peña, "Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data," *Sci. Rep.*, vol. 15, no. 1, p. 29420, Aug. 2025, doi: 10.1038/s41598-025-15388-9.

[37] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Min.*, vol. 16, no. 1, p. 4, Feb. 2023, doi: 10.1186/s13040-023-00322-4.

[38] J. Miao and W. Zhu, "Precision–recall curve (PRC) classification trees," *Evol. Intell.*, vol. 15, no. 3, pp. 1545–1569, Sep. 2022, doi: 10.1007/s12065-021-00565-2.

[39] R. Meléndez, R. Giraldo, and V. Leiva, "Sign, Wilcoxon and Mann-Whitney Tests for Functional Data: An Approach Based

on Random Projections," *Mathematics*, vol. 9, no. 1, p. 44, Dec. 2020, doi: 10.3390/math9010044.

[40] M. Q. R. Pembury Smith and G. D. Ruxton, "Effective use of the McNemar test," *Behav. Ecol. Sociobiol.*, vol. 74, no. 11, p. 133, Nov. 2020, doi: 10.1007/s00265-020-02916-y.

[41] Y. Atilkan *et al.*, "Advancing Crayfish Disease Detection: A Comparative Study of Deep Learning and Canonical Machine Learning Techniques," *Appl. Sci.*, vol. 14, no. 14, p. 6211, Jul. 2024, doi: 10.3390/app14146211.