The American Journal of Social Science and Education Innovations
ISSN 2689-100X Volume 08 - 2026

An Analytical Study Of Hybrid Artificial Intelligence Techniques For Multi-Modal Fake News Detection

Asaad Makki Al-Ebadi
Higher Institute for Security and Administrative Development, Ministry of Interior, Iraq

Aqeel Luaibi Challoob
Department of Computer Techniques Engineering, Imam Alkadhim University College (IKU), Baghdad, 10001, Iraq

Received: 03 Oct 2025 | Received Revised Version: 12 Oct 2025 | Accepted: 23 Oct2026 | Published: 02 Nov 2026

Volume 08 Issue 011 2026 | DOI: 10.37547/tajssei/Volume08Issue11-02

Abstract

Overview: The rapid growth of digital media has transformed news dissemination while enabling false information to spread widely, threatening information reliability and public trust. Automated fake news detection has consequently become an important research area, particularly where access is restricted to news headlines.

Research Problem: The research problem concerns accurately and reliably distinguishing fake from authentic news using headline content alone. Headlines provide limited contextual information and contain complex linguistic and semantic relationships that conventional statistical representations may inadequately capture. Additional challenges involve balancing predictive performance, generalizability, and computational cost while accommodating the changing nature of misinformation. These limitations motivate the investigation of complementary learning approaches within a consistent experimental framework.

Research Contributions: The study specifies a common headline-only framework for comparing SVM, XGBoost, CatBoost, BiLSTM, CNN–BiLSTM with Attention, RoBERTa, and DeBERTa. The proposed WELFake protocol assigns 60% of the stated 72,134 records to training, 20% to validation, and 20% to testing, with model-specific statistical, sequential, or contextual representations. A separate numerical illustration uses 10,000 generated observations, balanced between fake and authentic labels, to examine confusion matrices, accuracy, precision, recall, F1-score, and ROC–AUC. The illustration is not an evaluation of trained models on the WELFake test partition.

Research Conclusion: Within the generated-score analysis, the DeBERTa-labelled scenario records 94.78% accuracy, 94.79% F1-score, and 0.9889 ROC–AUC. These values follow the specified score distributions and do not establish empirical architectural superiority. Comparisons between SVM and CatBoost demonstrate that threshold-specific accuracy, positive-prediction reliability, and ranking discrimination can favour different scenarios. The study contributes a comparative protocol and a transparent interpretation of evaluation trade-offs. Verified held-out predictions, reproducible training configurations, and external evaluation are required before conclusions about WELFake performance, generalization, or deployment can be drawn.

Keywords: Fake News Detection; News Headline Classification; Machine Learning; Deep Learning; Transformer Models.

© 2026 Asaad Makki Al-Ebadi, Aqeel Luaibi Challoob.This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). The authors retain copyright and allow others to share, adapt, or redistribute the work with proper attribution.

Cite This Article: Asaad Makki Al-Ebadi, & Aqeel Luaibi Challoob. (2026). An Analytical Study Of Hybrid Artificial Intelligence Techniques For Multi-Modal Fake News Detection . The American Journal of Social Science and Education Innovations, 8(11), 12–28. https://doi.org/10.37547/tajssei/Volume08Issue11-02

1. Introduction

The rapid development of digital media has transformed how news is produced, distributed, and consumed. Online platforms enable information to reach large audiences quickly while creating opportunities for false and misleading content to circulate without adequate verification. Consequently, automated fake news detection has become an important research area involving machine learning and deep learning techniques that distinguish authentic from fabricated news using patterns learned from labelled examples [1], [4].

Existing research investigates conventional classifiers, neural networks, and hybrid architectures. Comparative studies examine differences between machine learning and deep learning approaches, while hybrid methods combine complementary mechanisms for textual feature extraction and classification [4], [5]. Contextual language models have also been investigated for fake news detection, including BERT-based approaches that represent words in relation to their surrounding text [6].

Within this research area, headline-based detection addresses situations in which only the news title is available. The present study proposes a comparative framework involving Support Vector Machine (SVM), XGBoost, CatBoost, Bidirectional Long Short-Term Memory (BiLSTM), CNN–BiLSTM with Attention, RoBERTa, and DeBERTa. These models represent conventional machine learning, sequential and hybrid deep learning, and Transformer-based approaches. Their inclusion supports an investigation of how different representations and learning mechanisms address the limited textual information available in news headlines.

The scale and changing nature of online information create difficulties for manual verification and rule-based filtering. Manual assessment requires substantial effort, whereas fixed rules may struggle to accommodate changes in the language and presentation of misinformation. These challenges motivate automated approaches that process textual information and identify patterns associated with fake and authentic news [7],[8].

Headline-based classification introduces additional difficulties because headlines provide less contextual information than complete articles. A headline may contain ambiguous wording, emotionally charged expressions, or compressed claims whose interpretation depends on information outside the title. The headline-focused research discussed in the literature therefore motivates investigating methods capable of identifying useful linguistic and semantic patterns within short texts [9].

Text representation is central to this problem. Statistical representations, such as term frequency–inverse document frequency (TF–IDF), describe lexical occurrence patterns but provide limited information about broader contextual relationships. Neural representations introduce mechanisms for learning sequential dependencies, while contextual models represent words according to their surrounding linguistic content. Previous research emphasizes that the suitability of a representation depends on the learning algorithm and the characteristics of the classification task [7], [10].

The proposed framework uses WELFake, described as containing 72,134 labelled news records with headline, article-text, and binary-label fields [31]. Only headlines and their associated labels are selected. This restriction establishes a common input scope for comparing the seven approaches while excluding article bodies, images, source information, and other metadata.

A further concern is the interpretation of evaluation results. Performance within a single dataset does not establish equivalent effectiveness across different sources, topics, languages, or domains. Generalization, interpretability, and computational requirements must therefore be considered alongside classification metrics when assessing the practical relevance of headline-based detection [11].

The central research problem is how to distinguish fake from authentic news accurately and reliably using headline content alone. Headlines provide limited contextual evidence and contain linguistic and semantic relationships that may be inadequately represented by conventional statistical features. Although sequential, hybrid, and Transformer-based approaches offer different mechanisms for representing this information, their relative

effectiveness requires evaluation under consistent data and assessment conditions [9], [10].

The methodological gap addressed by the proposed framework is the need to compare complementary model families within a unified headline-only setting. Differences in data selection, preprocessing, textual representations, and evaluation procedures can complicate the interpretation of model comparisons. A common protocol is therefore specified to examine SVM, XGBoost, CatBoost, BiLSTM, CNN–BiLSTM with Attention, RoBERTa, and DeBERTa using precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC–ROC).

Accordingly, the investigation is guided by three research questions:

How effectively can conventional machine learning approaches distinguish fake from authentic news using headlines alone?

To what extent can sequential, hybrid, and Transformer-based approaches improve the representation of linguistic, semantic, and contextual patterns in headlines?

Which investigated approach offers the most appropriate balance between classification performance, generalization, and computational efficiency?

The research questions motivate the proposed empirical comparison. The numerical evidence currently available addresses a narrower analytical question: how assumed score distributions produce different error profiles and evaluation metrics. The resulting comparisons are reported as generated-score scenarios, with their interpretation separated from claims about trained WELFake classifiers.

This study aims to develop a unified methodological framework for investigating headline-based fake news detection through a comparison of conventional machine learning, deep learning, hybrid neural, and Transformer-based approaches. The framework specifies model-appropriate representations: TF–IDF unigram–bigram features for SVM, XGBoost, and CatBoost; ordered token sequences for BiLSTM and CNN–BiLSTM with Attention; and architecture-specific subword representations for RoBERTa and DeBERTa. It also specifies stratified training, validation, and test partitions of 60%, 20%, and 20%, respectively, with validation used for configuration selection and the independent test subset reserved for final evaluation.

The intended comparison uses precision, recall, F1-score, and AUC–ROC to examine complementary aspects of classification. Within this analytical framework, the study also clarifies the interpretation of these measures through synthetic scenarios while distinguishing illustrative numerical outcomes from empirical findings.

1.1 Research Contributions

Building on the comparative and headline-focused research discussed in the literature [4], [9], the study provides the following methodological and analytical contributions:

1. Unified headline-based classification framework. The study specifies a common framework for investigating fake news detection using news headlines and their associated binary labels. Restricting the input to headlines establishes a consistent scope for comparing approaches without incorporating article bodies or additional metadata.

Comparison across seven classification approaches. The framework brings together SVM, XGBoost, CatBoost, BiLSTM, CNN–BiLSTM with Attention, RoBERTa, and DeBERTa, extending the proposed comparison across conventional machine learning, sequential deep learning, hybrid neural architectures, and pretrained Transformers.

Architecture-specific text representations. The methodology assigns TF–IDF unigram–bigram features to conventional classifiers, ordered token sequences to neural models, and model-specific subword inputs to Transformers. This design reflects the importance of matching textual representations to the learning approach emphasized in the reviewed literature [7].

Integration of complementary neural mechanisms. The proposed CNN–BiLSTM with Attention architecture combines local feature extraction, bidirectional sequence modelling, and weighted aggregation of contextual representations. Its inclusion provides a basis for investigating whether these mechanisms offer advantages over a standalone BiLSTM, without treating their individual contributions as empirically established.

Investigation of contextual headline representations. The framework includes RoBERTa and DeBERTa to examine contextual modelling under the limited information available in headlines. This investigation is motivated by the contextual and headline-classification approaches discussed in the literature [6], [9].

Consistent evaluation design. The proposed protocol specifies common stratified training, validation, and test partitions, with validation-based configuration selection and independent final testing. Precision, recall, F1-score, and AUC–ROC provide complementary criteria for

assessing prediction reliability, detection coverage, and discrimination.

Synthetic illustration of evaluation trade-offs. The numerical analysis demonstrates how confusion-matrix counts and evaluation metrics describe different aspects of classification behaviour. Because these results arise from constructed labels and score distributions, this contribution concerns metric interpretation rather than verified WELFake performance.

Together, these contributions establish a comparative methodological foundation. Empirical model superiority, improved generalization, computational savings, and multimodal effectiveness remain subjects for further validation.

This paper is organized as follows: Section 2 presents the proposed methodology, Section 3 presents the results and discussion, and Section 4 concludes the study, identifies its limitations, and outlines future research directions.

2. THE PROPOSED METHODOLOGY

2.1. Study Design and Scope

The study specifies a comparative framework for binary fake news classification from news headlines. The design examines three families of learning approaches under a common input scope: conventional machine learning, sequential and hybrid deep learning, and pretrained Transformers. This selection is motivated by the comparative, hybrid, and headline-focused approaches reviewed in the literature [4], [5], [9]. The seven candidate classifiers are Support Vector Machine (SVM), XGBoost, CatBoost, Bidirectional Long Short-Term Memory (BiLSTM), CNN–BiLSTM with Attention, RoBERTa, and DeBERTa.

The framework comprises headline and label selection, stratified partitioning, text preprocessing, model-specific representation, classifier development, validation-based configuration selection, and independent test evaluation. Two analyses are distinguished: a proposed empirical protocol using WELFake and a numerical illustration based on generated labels and continuous decision scores. The latter provides a controlled setting for interpreting evaluation measures. It does not use WELFake observations or establish measured performance for the named model architectures.

2.2. Data Source and Partitioning

WELFake is the designated source for the proposed empirical protocol. It is described as containing 72,134 labelled news records with title, article-text, and binary-label fields and a nearly balanced distribution of fake and authentic news [31]. Only the title and its associated label are selected. Article bodies, source and author information, other metadata, and non-textual modalities are excluded. The unit of analysis is consequently a headline–label pair, rather than a complete article or multimodal news item.The partitioning design allocates 60% of observations to training, 20% to validation, and 20% to testing using stratified sampling. The corresponding allocation is shown in Table 1. Stratification aims to preserve approximate class proportions; it does not establish independence between related headlines. The available reporting does not establish the retained counts following eligibility checks or document record-level overlap analysis.

Table 1. Data allocation specified for the proposed WELFake protocol.

Subset	Allocation	Documented count	Role
Training	60%	43,280	Parameter estimation
Validation	20%	14,427	Configuration selection
Testing	20%	14,427	Final held-out evaluation
Total	100%	72,134	Original stated allocation

Table 1 specifies the intended allocation before eligibility checks and does not establish verified post-cleaning sample sizes. The proposed protocol requires identical partition identities across classifiers, training-only fitting of representations, validation-based configuration selection, and final testing after model selection is complete. The stated WELFake test allocation contains 14,427 records, whereas the numerical analysis in Section 2.7 contains 10,000 generated observations. These populations must not be treated as interchangeable, and

the reported numerical results are not estimates from the proposed WELFake test partition.

2.3. Headline Preprocessing

The preprocessing pipeline standardizes headlines before numerical representation. Text normalization, cleaning, and feature extraction form part of the natural language processing workflow discussed in the reviewed literature [12], [13]. The specified sequence is lowercasing; removal of URLs, user mentions, hashtags, and HTML tags; removal of punctuation and digits; whitespace normalization; removal of common English stop words; and WordNet lemmatization.

The cleaned headlines are then passed to the appropriate representation branch. This common cleaning sequence defines the proposed procedure. Its effect on predictive performance, including its suitability for pretrained tokenizers, is not isolated by an ablation experiment in the available analysis.

2.4. Model-Specific Input Representations

2.4.1. Statistical Representation

SVM, XGBoost, and CatBoost receive sparse term frequency–inverse document frequency (TF–IDF) features. The representation includes unigrams and bigrams, preserving individual lexical items and consecutive word pairs. The importance of matching feature representations to learning algorithms motivates this branch of the design [7], [13]. For a term t in headline d, the weighting is expressed as:

$$\mathrm{TFIDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t). \qquad (1)$$

Here, $\mathrm{TF}(t,d)$ denotes term frequency within the headline and $\mathrm{IDF}(t)$ denotes inverse document frequency across the fitting corpus. The vectorizer is fitted exclusively on training headlines and reused, without refitting, to transform validation and test headlines.

2.4.2. Sequential Representation

BiLSTM and CNN–BiLSTM with Attention receive ordered word-index sequences. A vocabulary is constructed from training headlines, and each token is assigned an integer identifier. Padding and truncation standardize sequence lengths. An embedding layer maps the identifiers to dense vectors while preserving sequence order. This representation supports the embedding-based and bidirectional modelling approaches discussed in the reviewed literature [14], [15].

2.4.3. Transformer Representation

RoBERTa and DeBERTa receive subword token identifiers produced by their respective pretrained tokenizers, together with attention masks distinguishing valid tokens from padding. Padding and truncation standardize the inputs. These architecture-specific inputs are passed to the corresponding encoder and classification head [16], [17]. The proposed protocol does not supply exact checkpoint identifiers, maximum sequence lengths, or final tokenizer configurations.

2.5. Classifier Architecture and Model Selection

2.5.1. Conventional Machine Learning

The conventional branch compares a maximum-margin classifier with two gradient-boosting classifiers. SVM uses the sparse TF–IDF representation to construct a separating decision boundary [18]. XGBoost builds an ensemble through successive decision-tree additions with regularization [19]. CatBoost supplies an alternative boosting strategy [20]. All three receive the same type of numerical input; CatBoost's categorical-feature capabilities are therefore not an explanatory component of this design.

The methodology assigns regularization and other SVM settings, and the tree, learning-rate, and regularization settings of the boosting models, to validation-based configuration selection. No final settings are supplied in the available implementation description.

2.5.2. BiLSTM and the Hybrid Attention Model

BiLSTM processes an embedded headline in forward and backward directions, concatenating the directional hidden states to represent sequential context [15]. The resulting representation is passed to a classification layer. The CNN–BiLSTM with Attention architecture adds convolutional extraction of local patterns before bidirectional sequence modelling and applies attention-based aggregation before classification [21]. Its processing order is embedding, CNN, BiLSTM, attention, and classification.

For contextual hidden state $\mathbf{h}_t$, attention is specified as:

$$e_t = \mathbf{v}_a^T \tanh(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}_a), \qquad \alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{n} \exp\left(e_j\right)}. \qquad (2)$$

The aggregated headline representation is:

$$\mathbf{r} = \sum_{t=1}^{n} \alpha_t \, \mathbf{h}_t. \qquad (3)$$

Here, $\mathbf{W}_a$, $\mathbf{b}_a$, and $\mathbf{v}_a$ are attention parameters; e_t is the attention score at position t; α_t is its normalized weight; and n is the number of sequence positions. A classification layer maps $\mathbf{r}$ to the prediction. These operations specify architectural roles without establishing the empirical contribution of individual components.

The neural classification objective is binary cross-entropy:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]. \qquad (4)$$

In Equation (4), N is the number of observations, y_i is the actual binary label, and p_i is the predicted probability of the positive class.

2.5.3. Transformer Fine-Tuning

RoBERTa and DeBERTa provide pretrained contextual encoders followed by task-specific classification heads [16], [17]. DeBERTa is distinguished in the documented design by its separate treatment of token content and relative-position information within attention. Each encoder produces contextual hidden representations, and a sequence-level representation $\mathbf{h}_{\mathrm{cls}}$ is passed to the classifier:

$$\mathbf{z} = \mathbf{W}\mathbf{h}_{\mathrm{cls}} + \mathbf{b}, \qquad P(y \mid \mathbf{X}) = \mathrm{Softmax}(\mathbf{z}). \qquad (5)$$

Here, $\mathbf{X}$ denotes the tokenized headline, $\mathbf{W}$ and $\mathbf{b}$ are classification-head parameters, and $\mathbf{z}$ denotes class logits. The proposed procedure jointly fine-tunes the encoder and classification head using training data, with validation performance guiding configuration selection.

2.5.4. Configuration Selection and Test Isolation

The documented procedure fits each candidate using training data, assesses candidate configurations on validation data, retains the preferred configuration, and evaluates it once on the held-out test subset. Training-derived vectorizers and vocabularies remain fixed when processing validation and test inputs. This separation is specified to prevent test observations from influencing representation fitting or configuration selection.

Final search spaces, selected hyperparameters, optimization settings, neural dimensions, stopping criteria, and implementation environment details are not supplied. The protocol therefore describes the intended comparison; it does not establish a reproducible record of completed WELFake training.

2.6. Evaluation Protocol

Evaluation combines confusion-matrix analysis with threshold-based metrics and ROC–AUC. The use of complementary assessment criteria is consistent with the evaluation principles discussed in the reviewed literature [22]. Fake news is the positive class. A true positive (TP) is a fake observation classified as fake; a true negative (TN) is an authentic observation classified as authentic; a false positive (FP) is an authentic observation classified as fake; and a false negative (FN) is a fake observation classified as authentic.

Accuracy measures the proportion of correct classifications [23],[24],[25]:

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \qquad (6)$$

Precision and recall measure positive-prediction reliability and detection coverage, respectively:

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \qquad \mathrm{Recall} = \frac{TP}{TP + FN}. \qquad (7)$$

Their harmonic mean is:

$$F_1 = \frac{2TP}{2TP + FP + FN}. \qquad (8)$$

Precision, recall, and F1 are calculated for the positive class, rather than as macro or weighted averages. Threshold-based measures are multiplied by 100 when expressed as percentages. ROC analysis uses continuous scores to relate the true-positive rate to the false-positive rate across thresholds:

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}. \qquad (9)$$

AUC [26],[27] summarizes discrimination across the ROC curve. It is interpreted alongside the confusion matrix and threshold-based measures. The proposed WELFake protocol does not specify an empirically selected classification threshold; the zero threshold below belongs exclusively to the synthetic construction.

2.7. Synthetic Score-Based Analysis

The numerical illustration uses 10,000 constructed observations with 5,000 fake labels and 5,000 authentic labels. The same synthetic labels support all seven model-labelled scenarios. Score generation is specified using NumPy's random generator with seed 42. Conditional score distributions are:

$$s \mid y = 1 \sim \mathcal{N}\left(\frac{d}{2} + b, 1\right), \qquad s \mid y = 0 \sim \mathcal{N}\left(-\frac{d}{2} + b, 1\right). \qquad (10)$$

Here, $y = 1$ denotes fake news, $y = 0$ denotes authentic news, d is the assumed difference between the class-conditional score distributions, and b is a location shift relative to the decision threshold. A score greater than or equal to zero is classified as fake. Table 2 lists the scenario parameters.

Table 2. Distribution parameters for the synthetic model-labelled scenarios.

Scenario label	d	b
SVM	2.50	−0.03
XGBoost	2.35	0.05
CatBoost	2.45	0.02
BiLSTM	2.55	0.08
CNN–BiLSTM with Attention	2.85	0.06
RoBERTa	3.15	−0.02
DeBERTa	3.30	0.02

Threshold-based metrics use the generated class assignments, whereas ROC–AUC uses the continuous scores. The parameters in Table 2 are assumptions of the construction, not training hyperparameters, fitted model parameters, or estimates of architectural capabilities. The analysis illustrates metric behaviour and cannot establish a ranking of classifiers trained on WELFake. Its numerical outcomes are presented in Section 3.

3. RESULTS AND DISCUSSION

This section presents the numerical findings corresponding to the synthetic score-based analysis specified in Section 2.7. The seven labels identify constructed scenarios associated with the proposed classifier families. They do not identify verified predictions from models trained and tested on WELFake. Consequently, the findings support analysis of evaluation behaviour and error trade-offs, while empirical architectural superiority remains unestablished.

3.1 Evaluation population and evidence consistency

The analysis comprises 10,000 constructed observations, equally divided between fake news and authentic news. Fake news is the positive class (y = 1), and authentic news is the negative class (y = 0). Each scenario uses the same label population, continuous decision scores, and a zero classification threshold: scores greater than or equal to zero are classified as fake. The distribution parameters and reported random seed are specified in Section 2.7. These assumptions govern the illustrative outcomes and are separate from the proposed 60:20:20 WELFake partition.

Accuracy, precision, recall, and F1-score were verified by recalculation from the confusion-matrix counts. The reported percentages agree with these counts after rounding to two decimal places, and every scenario contains 5,000 observations per actual class. Precision, recall, and F1-score refer to the positive class rather than macro or weighted averages. ROC–AUC[28],[29],[30] values summarize continuous-score discrimination, whereas the confusion matrices describe classification at the zero threshold. Observation-level score arrays are unavailable for independent recalculation of AUC. Differences between percentages are expressed in percentage points.

3.2 Overall performance and classification errors

Table 3. Overall threshold-based performance across the seven synthetic scenarios

Model	Model family	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)
SVM	Machine learning	89.10	89.49	89.05	88.60
XGBoost	Machine learning	87.13	85.78	87.37	89.02
CatBoost	Machine learning	89.20	88.63	89.28	89.94
BiLSTM	Deep learning	89.86	88.71	90.01	91.34
CNN–BiLSTM with Attention	Hybrid deep learning	92.60	92.13	92.64	93.16
RoBERTa	Transformer	94.07	94.31	94.05	93.80
DeBERTa	Transformer	94.78	94.62	94.79	94.96

Table 3 shows that the DeBERTa-labelled scenario has the largest accuracy (94.78%), precision (94.62%), F1-score (94.79%), and recall (94.96%). RoBERTa and CNN–BiLSTM with Attention follow across these measures. Among conventional-model scenarios, CatBoost has the highest accuracy, F1-score, and recall, whereas SVM has the highest precision. This ordering describes the assumed score distributions; it cannot establish that the corresponding trained architectures would exhibit the same ranking.

Table 4. Supporting confusion-matrix counts

Model	TP	TN	FP	FN	Total
SVM	4,430	4,480	520	570	10,000
XGBoost	4,451	4,262	738	549	10,000
CatBoost	4,497	4,423	577	503	10,000
BiLSTM	4,567	4,419	581	433	10,000
CNN–BiLSTM with Attention	4,658	4,602	398	342	10,000
RoBERTa	4,690	4,717	283	310	10,000
DeBERTa	4,748	4,730	270	252	10,000

In Table 4, true positives are fake observations correctly detected, true negatives are authentic observations correctly retained, false positives are authentic observations flagged as fake, and false negatives are fake observations missed. Figure 1 presents the associated matrices. Their common population and shared colour scale permit direct visual comparison of the two error types.

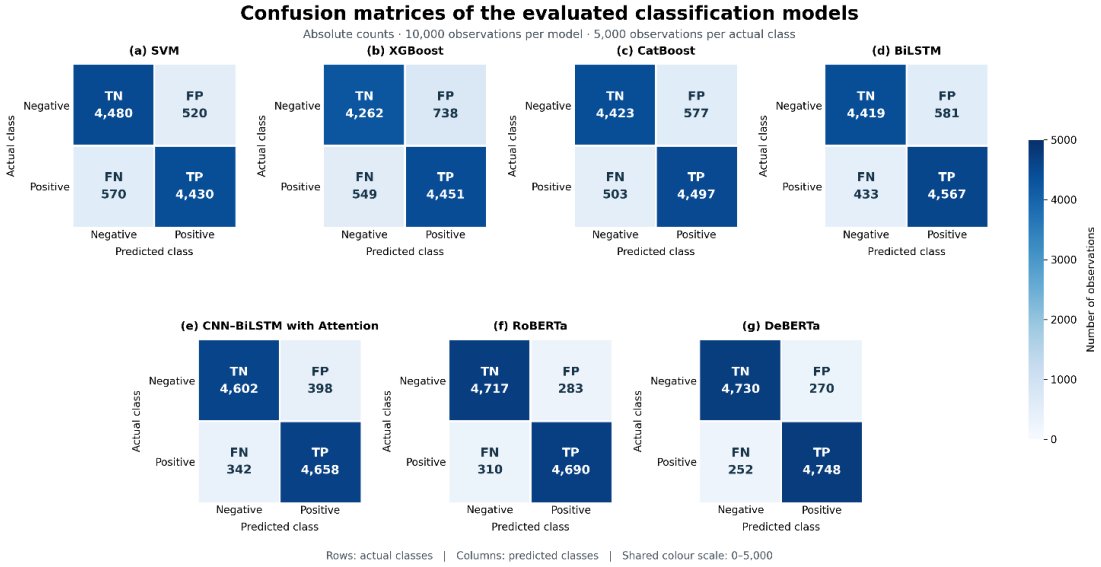


Figure 1. Confusion matrices for the seven synthetic classification scenarios shown in panels (a)–(g). Rows

represent actual classes and columns represent predicted classes; positive denotes fake news and negative denotes authentic news.

The SVM and CatBoost comparison demonstrates why aggregate accuracy is insufficient. CatBoost detects 67 additional fake observations (4,497 versus 4,430) but incorrectly flags 57 additional authentic observations (577 versus 520). It therefore produces only ten fewer errors overall. BiLSTM similarly detects 70 more fake observations than CatBoost while adding four false positives. These error exchanges are relevant when screening objectives assign different costs to missed misinformation and incorrectly restricted authentic content.

The CNN–BiLSTM with Attention scenario reduces false positives from 581 to 398 and false negatives from 433 to 342 relative to BiLSTM, yielding 274 fewer errors. RoBERTa reduces the combined error count further to 593, and DeBERTa to 522. These are descriptive comparisons within the constructed population. They do not isolate benefits attributable to convolution, bidirectional processing, attention, or pretrained contextual representations.

3.3 Accuracy and total classification error

Table 5. Correct and incorrect classifications underlying accuracy

Model	Correct classifications	Incorrect classifications	Accuracy (%)
SVM	8,910	1,090	89.10
XGBoost	8,713	1,287	87.13
CatBoost	8,920	1,080	89.20
BiLSTM	8,986	1,014	89.86
CNN–BiLSTM with Attention	9,260	740	92.60
RoBERTa	9,407	593	94.07
DeBERTa	9,478	522	94.78

Accuracy ranges from 87.13% for XGBoost to 94.78% for DeBERTa (Table 5), a difference of 7.65 percentage points and 765 correct classifications. RoBERTa reaches 94.07%, while CNN–BiLSTM with Attention reaches

92.60%. Relative to BiLSTM, the hybrid scenario increases accuracy by 2.74 percentage points; the DeBERTa–RoBERTa difference is 0.71 percentage points, corresponding to 71 classifications.

Because both actual classes contain 5,000 observations, each class contributes equally to accuracy. Nevertheless, accuracy conceals how errors are distributed, as the small 0.10-percentage-point difference between SVM and CatBoost illustrates. The reported gaps provide no evidence of statistical significance or stability across training runs. No repeated empirical fits, uncertainty estimates, or paired significance analysis are supplied, so the numerical ordering should remain descriptive.

3.4 Precision and false-positive burden

Table 6. Correct fake predictions and false positives underlying precision

Model	Correct fake predictions	Real observations flagged as fake	Precision (%)
SVM	4,430	520	89.49
XGBoost	4,451	738	85.78
CatBoost	4,497	577	88.63
BiLSTM	4,567	581	88.71
CNN–BiLSTM with Attention	4,658	398	92.13
RoBERTa	4,690	283	94.31
DeBERTa	4,748	270	94.62

Table 6 relates precision to the reliability of fake-news predictions. DeBERTa achieves 94.62% precision from 4,748 true positives and 270 false positives. XGBoost has the lowest precision, 85.78%, together with the largest false-positive count, 738. SVM achieves 89.49% precision compared with 88.63% for CatBoost, despite detecting fewer fake observations. Thus, improving detection coverage does not necessarily improve the reliability of positive predictions.

False positives can increase human review requirements and incorrectly challenge authentic material. Precision, however, is conditional on both the decision rule and class prevalence. The balanced construction does not represent every deployment population: at unchanged sensitivity and specificity, a lower prevalence of fake news would reduce positive predictive value. Accordingly, the reported precision values should not be transferred directly to an operational news stream.

3.5 Joint interpretation of precision recall and F1 score

Table 7. Positive-class precision recall and F1-score

Model	Precision (%)	Recall (%)	F1-score (%)
SVM	89.49	88.60	89.05
XGBoost	85.78	89.02	87.37
CatBoost	88.63	89.94	89.28
BiLSTM	88.71	91.34	90.01
CNN–BiLSTM with Attention	92.13	93.16	92.64
RoBERTa	94.31	93.80	94.05
DeBERTa	94.62	94.96	94.79

The F1-scores in Table 7 range from 87.37% for XGBoost to 94.79% for DeBERTa. BiLSTM combines 88.71% precision and 91.34% recall to obtain an F1-score of 90.01%. CNN–BiLSTM with Attention combines 92.13% precision and 93.16% recall to obtain 92.64%, exceeding BiLSTM by 2.63 percentage points. RoBERTa exceeds the hybrid scenario by 1.41 percentage points, while DeBERTa exceeds RoBERTa by 0.74 percentage points.

The harmonic mean limits the extent to which one strong component compensates for a weaker one. XGBoost illustrates this relationship because its recall of 89.02% is accompanied by lower precision of 85.78%. F1 summarizes the positive-class trade-off but does not directly incorporate true negatives or specify an application-specific error cost. It should therefore be considered alongside the full confusion matrix rather than treated as a complete measure of utility.

3.6 Recall and missed fake observation

Table 8. Detected and missed fake observations underlying recall

Model	Fake observations detected	Fake observations missed	Recall (%)
SVM	4,430	570	88.60
XGBoost	4,451	549	89.02
CatBoost	4,497	503	89.94
BiLSTM	4,567	433	91.34
CNN–BiLSTM with Attention	4,658	342	93.16
RoBERTa	4,690	310	93.80
DeBERTa	4,748	252	94.96

Table 8 shows that DeBERTa detects 4,748 of the 5,000 fake observations, giving 94.96% recall and 252 false negatives. SVM detects 4,430, giving 88.60% recall and 570 false negatives. The difference amounts to 318 additional detections and 6.36 percentage points. RoBERTa and CNN–BiLSTM with Attention miss 310 and 342 fake observations, respectively.

Higher recall can support preliminary screening when flagged material undergoes subsequent human assessment. However, XGBoost detects only 21 more

fake observations than SVM while incorrectly flagging 218 additional authentic observations. Its 0.42-percentage-point recall advantage must therefore be interpreted with the accompanying false-positive burden.

The zero threshold is an assumption of the construction, not a validated operating point selected for a trained classifier. Threshold selection for deployment remains unresolved.

3.7 ROC curves and discrimination across thresholds

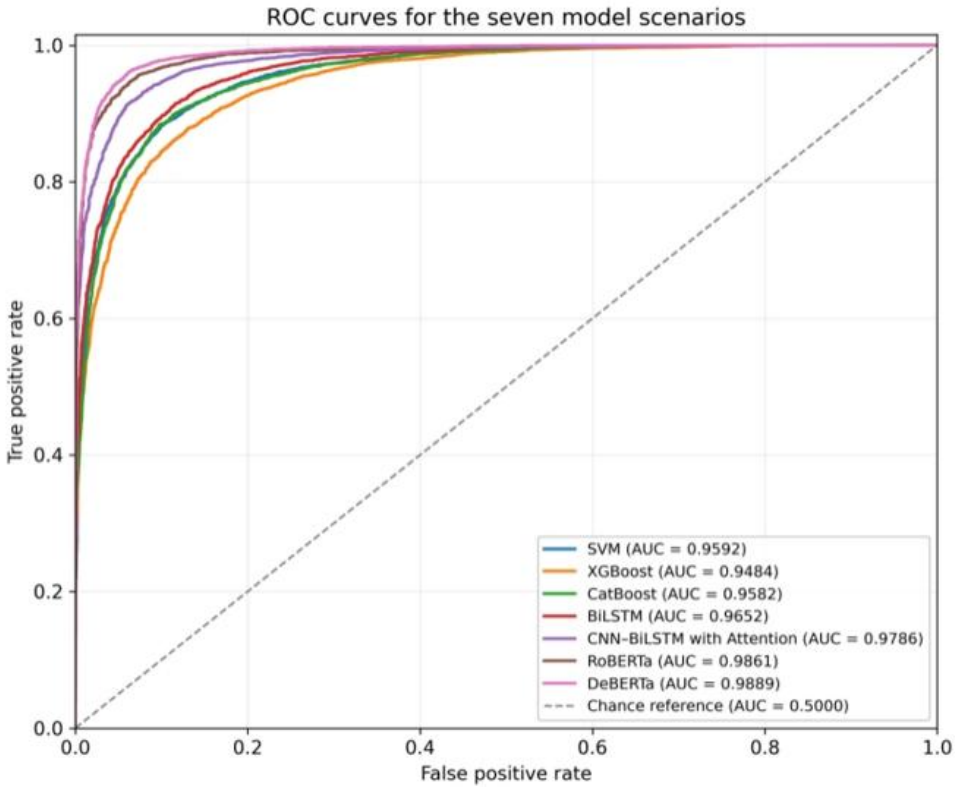


Figure 2. ROC curves and corresponding AUC values for the seven model

Figure 2 extends the comparison beyond the zero threshold. The reported AUC values are 0.9889 for DeBERTa, 0.9861 for RoBERTa, 0.9786 for CNN–BiLSTM with Attention, 0.9652 for BiLSTM, 0.9592 for SVM, 0.9582 for CatBoost, and 0.9484 for XGBoost. The chance reference has AUC = 0.5000. The figure and its legend preserve all seven reported curves and values.

The contrast between SVM and CatBoost separates ranking quality from threshold-specific classification: SVM has slightly higher AUC despite lower accuracy at the selected threshold. AUC summarizes how scores rank positive relative to negative observations; it does not establish factual understanding, calibration, or performance at every operationally relevant threshold. Where false alarms carry substantial costs, recall at a specified low false-positive rate would provide complementary evidence, but that quantity is not reported and is not inferred here.

3.8 Interpretation in relation to the research questions

For the first research question, the conventional scenarios illustrate that different metrics can favour different operating characteristics. CatBoost has slightly higher accuracy and F1 than SVM, whereas SVM has higher precision and AUC. This provides a methodological example of evaluating lexical-feature classifiers through complementary criteria; it does not establish their effectiveness on actual headlines.

For the second research question, the sequential, hybrid, and Transformer-labelled scenarios have progressively stronger aggregate values in the presented construction.

The proposed methodology assigns these families different representations: ordered word sequences, combined convolutional and sequential processing with attention, and pretrained contextual subword encodings. Their potential benefits motivate the comparison, but the score distributions were assumed rather than learned from headlines. Consequently, the numerical findings cannot demonstrate improved linguistic representation or identify which architectural component causes a gain. A controlled empirical comparison and component-level ablations would be required for those conclusions.

For the third research question, the evidence cannot identify the approach offering the best balance of predictive performance, generalization, and computational efficiency. Training time, inference latency, memory usage, and cross-domain performance were not measured. The largest illustrative metric values therefore cannot justify selecting an architecture for deployment. Such selection requires verified predictions, operational error costs, external evaluation, and resource measurements under documented conditions.

The comparative machine learning and deep learning literature [4], hybrid modelling work [5], representation research [7], and headline-based classification research [9] provide methodological context for these questions. WELFake [31] supplies the designated dataset for the proposed empirical protocol. Direct numerical claims of improvement over these studies are unwarranted because the present outcomes are synthetic and the comparison conditions are not matched. The cited studies contextualize the design rather than validate its numerical ranking.

3.9 Validity boundaries and practical implications

The principal validity boundary is the separation between the proposed WELFake protocol and the constructed numerical population. The reported tables and figures do not provide held-out WELFake performance estimates. Headline-only classification also omits qualifications and evidence available in article bodies, so even verified predictive performance would not by itself establish factual verification. The present analysis concerns textual headlines exclusively; images and other modalities are outside its scope.

Stratification preserves approximate class proportions but does not establish independence between related records. The proposed evaluation does not include a verified analysis of duplicate or near-duplicate overlap, shared publication sources, or topic leakage across partitions. These are unresolved threats to an eventual empirical evaluation, rather than demonstrated causes of the synthetic outcomes. Generalization to other datasets, time periods, languages, and misinformation contexts remains untested.

Reproducibility is also limited by the absence of linked training artifacts, finalized model configurations, and observation-level prediction arrays. The reported random seed and distribution parameters do not establish complete reproduction without the generator implementation and draw order. No uncertainty intervals, repeated-run variability, calibration analysis, error-content inspection, or measured computational costs are available. These omissions restrict the inferences that can be drawn and identify the evidence needed before practical deployment.

Taken together, the results show how overall correctness, positive-prediction reliability, detection coverage, and ranking discrimination can yield different assessments of the same constructed classification scenarios. DeBERTa has the largest reported values across all five metrics, followed by RoBERTa and CNN–BiLSTM with Attention, within the assumed construction. The analysis clarifies the interpretation of complementary evaluation measures; the relative performance of trained fake news detectors remains an empirical question.

4. CONCLUSION, LIMITATIONS, AND FUTURE DIRECTIONS

4.1 Conclusion

This study establishes a methodological framework for comparing seven approaches to headline-based fake news detection: SVM, XGBoost, CatBoost, BiLSTM, CNN–BiLSTM with Attention, RoBERTa, and DeBERTa. The framework aligns statistical, sequential, and contextual text representations with the corresponding classifier families and specifies a common stratified training, validation, and testing protocol. Its contribution is a structured basis for comparative investigation, complemented by a synthetic analysis of classification metrics and error trade-offs.

Across the 10,000-observation synthetic population, the DeBERTa-labelled scenario records the highest accuracy (94.78%), precision (94.62%), recall (94.96%), F1-score (94.79%), and ROC–AUC (0.9889), followed by RoBERTa and CNN–BiLSTM with Attention. These values summarize behaviour under assumed score

distributions; they do not establish the relative performance of architectures trained on WELFake. The conventional-model scenarios further demonstrate that a single ranking can obscure meaningful differences: CatBoost has higher accuracy than SVM, whereas SVM has higher precision and AUC. Joint interpretation of the confusion matrix, threshold-based metrics, and ROC analysis therefore provides a more informative assessment than any isolated measure.

The analysis addresses the research questions at the level of evaluation design and interpretation. It clarifies how conventional-model scenarios exchange detection coverage for false-positive burden and how sequential, hybrid, and contextual approaches can be compared under a shared protocol. However, the effectiveness of these approaches on actual headlines, the contribution of individual architectural components, and the balance between predictive performance and computational efficiency remain unresolved empirically. Headline classification may support preliminary screening, but its predictions alone cannot establish the factual validity of a news claim. The study consequently provides an analytical foundation for subsequent verification rather than a validated basis for automated moderation.

4.2 Limitations

The principal limitation concerns the scope of inference supported by the comparative analysis. The scenario-based design illustrates relationships among evaluation metrics, but it does not isolate the effects of model architecture, representation, or optimization. Numerical differences should therefore be interpreted descriptively rather than as evidence of statistically significant improvements or architectural superiority. Without controlled ablation experiments, differences involving the hybrid model cannot be attributed specifically to convolution, bidirectional sequence processing, or attention.

The proposed empirical protocol uses English headlines from one designated dataset and excludes article bodies, source information, publication context, user interactions, and non-textual modalities. This scope omits evidence that may be necessary to interpret ambiguous or misleading titles. It also leaves cross-source, cross-domain, temporal, and multilingual generalization unassessed. The present analysis does not establish that a trained system would maintain its performance on Arabic news or on emerging misinformation topics, and it provides no evaluation of multimodal detection.

Dataset independence and implementation reproducibility are not established. Stratification specifies approximate class proportions but does not prevent related headlines or shared sources from occurring across partitions. Duplicate and near-duplicate overlap has not been verified, so leakage remains a potential threat to an eventual empirical evaluation rather than a demonstrated explanation of the reported results. Retained sample counts, partition identities, complete final configurations, and linked prediction artifacts are not available as a verified implementation record. Likewise, the synthetic random seed and distribution parameters require the exact generator implementation and draw order for full numerical reproduction.

The balanced synthetic population and fixed zero threshold restrict operational interpretation. Precision depends on class prevalence, while the relative costs of false positives and false negatives vary by application. No validation-selected operating threshold, probability calibration, or evaluation under changing prevalence is reported. Repeated-run variability, confidence intervals, and paired uncertainty analyses are also absent; small descriptive differences therefore cannot establish stable comparative advantages. These omissions limit conclusions about reliability in real news streams.

Practical utility remains unmeasured. Training time, inference latency, memory consumption, explanation fidelity, and human review outcomes have not been evaluated. Accordingly, headline-only inputs cannot be claimed to deliver measured computational savings, and higher illustrative scores cannot justify deployment of a more complex architecture. The framework also does not establish whether model explanations would identify evidentially relevant cues rather than dataset-specific language patterns.

4.3 Future Directions

The immediate priority is a reproducible empirical comparison on the designated headline dataset. This work should document eligibility criteria, exclusions, retained class counts, split identifiers, preprocessing, model checkpoints, search spaces, selected hyperparameters, software versions, and hardware. Exact duplicates and related headline groups should be identified before partitioning, with source-aware separation where suitable metadata are available. Common held-out observations, comparable tuning budgets, and training-only fitting of feature transformations would support defensible comparisons. Releasing evaluation code, prediction

scores, and executable synthetic-generation code would enable independent checking of both analyses.

Repeated training runs and uncertainty estimates should then assess whether observed differences persist beyond random initialization and sampling variation. Paired comparisons should use predictions on the same test observations, with procedures appropriate to each metric. Controlled ablations should separately evaluate convolution, bidirectional processing, attention, and preprocessing choices, including their compatibility with pretrained tokenizers. These experiments would distinguish benefits supported by empirical evidence from differences arising from model configuration or optimization.

External and time-ordered evaluations should assess transfer beyond a single collection. Independent datasets and later publication periods would help identify reliance on source, topic, or temporal patterns. Arabic and other language settings should be examined with appropriate tokenization and multilingual or language-specific encoders. Domain adaptation and continual learning could subsequently be investigated using chronological protocols that prevent access to future labels. Such evaluation would establish whether adaptation improves transfer while retaining previously learned capabilities.

Controlled input comparisons should test whether full article bodies, source and propagation information, or aligned visual content add value beyond headlines. Multimodal extensions should be evaluated as new configurations with explicit modality availability, matched partitions, and ablations of each added input. Ensemble approaches may also be investigated where verified models make complementary errors, using validation data to select combination rules and preserving the independent test set. These extensions would test incremental benefit rather than assume that additional information or complexity necessarily improves detection.

Operational studies should select thresholds on validation data according to explicit review capacity and error costs, then evaluate them on untouched test data. Calibration, precision–recall analysis, and sensitivity to class prevalence would clarify the reliability of model outputs in different screening environments. Computational benchmarking should measure training cost, inference latency, throughput, and memory under common hardware conditions. Explanation methods should be checked for fidelity and usefulness through error analysis and human assessment. A prospective screening pilot with human review would then be needed to examine practical benefit, false-alarm burden, and robustness before broader use.

BIBLIOGRAPHY



Asaad M. Al-Ebadi (Iraqi) holds a bachelor's degree in Computer Engineering, awarded in 2018, and currently serves as a Captain in the Iraqi Ministry of Interior (MoI). He joined the MoI in 2020 and was commissioned as a Second Lieutenant in 2021(Al-Shumokh Course – D 27), subsequently promoted to First Lieutenant in 2023 and Captain in July 2025. He is currently pursuing a postgraduate diploma at the Higher Institute for Security and Administrative Development, affiliated with the Iraqi MoI. He welcomes research collaborations in information security, artificial intelligence, image processing, and the application of information security techniques to crime classification.

References

1. M. M. Lawal and A. Abdulrauf, "Fake news detection using bi-lstm architecture: A deep learning approach on the isot dataset," *J. Comput. Theor. Appl.*, vol. 3, no. 2, pp. 104–114, 2025.
2. F. A. Alshuwaier and F. A. Alsulaiman, "Fake news detection using machine learning and deep learning algorithms: A comprehensive review and future perspectives," *Computers*, vol. 14, no. 9, p. 394, 2025.
3. 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, 2025.
4. J. Alghamdi, Y. Lin, and S. Luo, "A comparative study of machine learning and deep learning techniques for fake news detection," *Information*, vol. 13, no. 12, p. 576, 2022.
5. A. K. Yadav *et al.*, "Fake news detection using hybrid deep learning method," *SN Comput. Sci.*, vol. 4, no. 6, p. 845, 2023.
6. R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, 2021.
7. M. A. B. Al-Tarawneh, O. Al-Irr, K. S. Al-Maaitah, H. Kanj, and W. H. F. Aly, "Enhancing fake news

detection with word embedding: A machine learning and deep learning approach," *Computers*, vol. 13, no. 9, p. 239, 2024.

8. H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," *Front. big Data*, vol. 6, p. 1320800, 2024.

9. 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, 2025.

10. J. Alghamdi, Y. Lin, and S. Luo, "Machine learning and deep learning approaches for fake news detection and related topics in multilingual contexts: A systematic literature review," *Multimed. Tools Appl.*, vol. 85, no. 4, p. 353, 2026.

11. Y. Tian, S. Xu, Y. Cao, Z. Wang, and Z. Wei, "An empirical comparison of machine learning and deep learning models for automated fake news detection," *Mathematics*, vol. 13, no. 13, p. 2086, 2025.

12. S. Bird, E. Klein, and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit . " O'Reilly Media, Inc.," 2009.

13. A. Tabassum and R. R. Patil, "A survey on text pre-processing & feature extraction techniques in natural language processing," Int. Res. J. Eng. Technol. , vol. 7, no. 06, pp. 4864–4867, 2020.

14. S. Selva Birunda and R. Kanniga Devi, "A review on word embedding techniques for text classification," Innov. Data Commun. Technol. Appl. Proc. ICIDCA 2020 , pp. 267–281, 2021.

15. L. Abualigah, Y. Y. Al-Ajlouni, M. S. Daoud, M. Altalhi, and H. Migdady, "Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe," Soc. Netw. Anal. Min. , vol. 14, no. 1, p. 40, 2024.

16. A. A. Pal, S. Mondal, C. A. Kumar, and C. J. Kumar, "A transformer-based approach for fake news and spam detection in social media using roberta," in 2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI) , IEEE, 2025, pp. 1256–1263.

17. S. Sweidan, N. A. Farouk, M. Abouhawwash, S. S. Askar, and M. Taha, "DeBERTa-based framework for detecting machine-generated content on social media: a comparative study," J. Big Data , vol. 13, no. 1, p. 10, 2026.

18. M. Sudhakar and K. P. Kaliyamurthie, "Detection of fake news from social media using support vector machine learning algorithms," Meas. Sensors , vol. 32, p. 101028, 2024.

19. J. Subramanian, G. J. Sendur, P. Adithiyan, S. P. Kumar, S. M. Kumar, and M. R. Kumar, "Enhancing fake news detection: A comprehensive framework leveraging XGBoost and CNNs," in 2024 3rd International conference on artificial intelligence for internet of things (AIIoT) , IEEE, 2024, pp. 1–6.

20. U. Gupta, S. Raj, and J. B. Pernabas, "Optimized Fake News Detection in Social Networks Using Boosting Algorithms and Machine Learning Classifiers," in International Conference on Computing and Machine Learning , Springer, 2025, pp. 275–287.

21. J. Alghamdi, Y. Lin, and S. Luo, "Modeling fake news detection using bert-cnn-bilstm architecture," in 2022 IEEE 5th international conference on multimedia information processing and retrieval (MIPR) , IEEE, 2022, pp. 354–357.

22. P. Resnik and J. Lin, "11 evaluation of NLP systems," Handb. Comput. Linguist. Nat. Lang. Process. , vol. 57, 2010.

23. Sideek, Shirin Muataz Mohammed, et al. "An Improved Anomaly-based Intrusion Detection System for IoT Applications using Machine Learning Methods." Pertanika Journal of Science & Technology 34.1 (2026): 401.

24. Nuimi, Zeina N., et al. "Decoupling Transient from Steady-state Performance in Fractional-order PI Control of a DC Microgrid: A Bi-objective Formulation Validated Against Single-metric and Recent-metaheuristic Benchmarks." *Int. J. Intell. Eng. Syst* 19.9 (2026): 337.

25. Yaseen, Nabeel Abdulrazaq, et al. "Enhancing Road Safety through Advanced Machine Learning: A Comparative Study of Tree Ensembles, Neural Networks, and Integrated Models for Multi-class Accident Severity Prediction under Class Imbalance." *Int. J. Intell. Eng. Syst* 19.9 (2026): 467.

26. Ahmed, Asraa Safaa, et al. "An accurate model for text document classification using machine learning

techniques." *Ing. des Syst. d'Information* 30.4 (2025): 913-921.

27. Jabbar, Noor Kareem, Majn Naderan, and Mustafa Sabah Taha. "HybridIoMT: A Dual-Phase Machine Learning Framework for Robust Cybersecurity in Internet of Medical Things." *International Journal of Intelligent Engineering & Systems* 18.4 (2025): 307.

28. Ahmed, Attallah Salih, Israa Nasir Abood, and Mustafa Sabah Taha. "Adaptive Multi-objective Optimization for Obstacle-aware Wireless Sensor Network Deployment: A Comparative Analysis of State-of-the-Art Algorithms." *International Journal of Intelligent Engineering & Systems* 19.2 (2026): 707.

29. Ismael, Bahaa Muneer, et al. "Non-dominated sorting genetic algorithm for channel assignment in multiple radio interfaces with multiple channels." *AIP Conference Proceedings*. Vol. 3393. No. 1. AIP Publishing LLC, 2026.

30. Ismael, Bahaa Muneer, et al. "Multi-Agent Reinforcement Learning for User-Router Assignment in Multi-Radio Multi-Channel Wireless Mesh Networks." *International Journal of Intelligent Engineering & Systems* 18.8 (2025): 316.

31. P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word embedding over linguistic features for fake news detection," IEEE Trans. Comput. Soc. Syst. , vol. 8, no. 4, pp. 881–893, 2021.