IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2385∼2397
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2385-2397

Enhancing fake news detection: a hybrid BERT-XGBoost model for improved performance and interpretability

Nishant Vasantkumar Hegde[1], Suneesh Bare[1], Namruth Reddy[2], Rajat Gondkar Aravinda[1], Minal Moharir[1], Aamir Ibrahim[1]
[1]Department of Computer Science and Engineering, RV College of Engineering, Bengaluru, India
[2]Senior Security Engineer, NVIDIA Corporation, Santa Clara, United States

Article Info

Article history:

Received Sep 29, 2025
Revised Mar 6, 2026
Accepted Apr 22, 2026

Keywords:

BERT
Deep learning
Fake news detection
Machine learning
Model interpretability
Natural language processing
XGBoost

ABSTRACT

The widespread spread of fake news poses a serious threat to the integrity of information. The dominant approach to detection involves end-to-end fine-tuning of large transformer models like bidirectional encoder representations from transformers (BERT), which, despite achieving high accuracy, often function as opaque "black boxes" with limited interpretability. This paper proposes and validates a hybrid, decoupled architecture that proves to be a more practical and powerful alternative. We first fine-tune a DistilBERT model on the full WELFake dataset of 71,537 articles after cleaning to create domain-specific embeddings. These high-dimensional vectors are then used as input features to train a robust extreme gradient boosting (XGBoost) classifier. The results demonstrate that the hybrid model achieves a state-of-the-art accuracy of 99.76%, slightly surpassing the already high performance of a standard end-to-end fine-tuned model. Crucially, this approach provides this top-tier performance while offering significant advantages in model interpretability through feature importance analysis. This work establishes that a decoupled architecture is not just a viable alternative but a superior practical strategy for combating misinformation, successfully balancing state-of-the-art accuracy with essential model transparency.

Corresponding Author:

Nishant Vasantkumar Hegde
Department of Computer Science and Engineering, RV College of Engineering
Bengaluru, India
Email: hegde.nishant2005@gmail.com

1. INTRODUCTION

In today's digital era, the swift circulation of misinformation and fake news has become a major societal concern. Maliciously crafted content, designed to deceive and manipulate public opinion, can destabilize democratic processes, erode public trust, and cause significant harm [1], [2]. The increasing sophistication of such content, often leveraging sensationalism and emotional language, makes manual detection unfeasible at scale, necessitating the development of advanced automated systems [3]. This has spurred extensive research into computational techniques capable of discerning false information with high accuracy and reliability [4], [5].

Initial methods for detecting fake news depended on conventional machine learning models that leveraged basic linguistic features. Approaches using term frequency–inverse document frequency (TF-IDF)

and n-grams combined with classifiers such as support vector machines (SVMs) and naive Bayes established the foundational framework [6], [7]. However, these methods often struggle to capture the subtle contextual and semantic nuances that differentiate sophisticated fake news from legitimate reporting, as they treat text as a mere "bag of words" [8]. The advent of deep learning, particularly transformer-based models like bidirectional encoder representations from transformers (BERT), marked a paradigm shift [9]–[11]. As demonstrated by numerous studies [12], [13], models like BERT excel because of their pre-training on vast text corpora, which endows them with a deep understanding of language context, a critical capability where previous models failed.

The current state-of-the-art methodology typically involves taking a pre-trained BERT-like model and fine-tuning it end-to-end on a specific fake news dataset. This approach has proven highly effective, achieving impressive accuracy by adapting the model's millions of parameters to classification task [12], [13]. However, this performance comes at a significant cost: interpretability. A fine-tuned transformer operates as a "black box", making it nearly impossible to understand or explain why it classified a particular news article as fake [14]. This lack of transparency is a major barrier to trust and adoption in critical applications, where accountability and the ability to audit model decisions are paramount [15]. In fields like journalism and policy-making, a model's prediction is often insufficient without a corresponding explanation. Recent work has highlighted concerns about adversarial robustness [16] and the need for source credibility assessment [17] alongside content-based detection.

To address this critical trade-off between performance and transparency, this study proposes and investigates a hybrid, decoupled architecture that synergizes the strengths of both deep learning and classical machine learning. The contributions are threefold. First, demonstrate that decoupled hybrid architecture achieves a state-of-the-art accuracy of 99.76%, slightly surpassing the already high accuracy of a fully fine-tuned DistilBERT model. Second, establish that this top-tier performance is achieved without the "black-box" trade-off, providing full model interpretability through extreme gradient boosting (XGBoost)'s feature importance analysis, a crucial step towards explainable artificial intelligence (XAI) in this domain [18], [19]. Finally, make the case that the hybrid model is a more practical and trustworthy architecture for real-world deployment, as it delivers superior accuracy, transparency, and greater inference efficiency, representing a pragmatic and generalizable template for XAI in other high-stakes text classification tasks where transparency is crucial, such as hate speech detection, sentiment analysis in sensitive contexts, and the identification of medical misinformation.

2. RELATED WORK

The scholarly exploration of automated fake news detection has advanced considerably, shifting from conventional statistical techniques to advanced deep learning models. This evolution can be broadly categorized into several key phases, each building upon the last to address the increasing complexity of misinformation. These advancements highlight a continuous search for models that are not only accurate but also robust and understandable.

2.1. Early machine learning approaches

Initial research efforts focused on manual feature engineering, extracting lexical, syntactic, and content-based features from news articles. Models such as naive Bayes, logistic regression, and SVMs were frequently employed using features like TF-IDF vectors, word frequencies, and readability metrics [6], [7]. These approaches often incorporated a rich set of handcrafted features, including stylistic attributes like punctuation frequency and capitalization, as well as psychological features derived from sentiment analysis. While these models provided a strong baseline and were computationally efficient, their primary limitation was the reliance on surface-level features. This made them vulnerable to simple adversarial attacks (e.g., minor text alterations) and fundamentally unable to grasp deeper semantic meanings or contextual nuances, a core challenge identified in foundational surveys of the field [8].

2.2. Deep learning and sequential models

The advent of deep learning brought forth models that could automatically learn feature representations, thereby minimizing the reliance on extensive manual feature engineering. A significant step forward came with the application of convolutional neural networks (CNNs) for text classification, which proved effective at capturing local patterns and n-gram-like features from text. Subsequently, recurrent neural networks (RNNs) and their advanced variants, long short-term memory (LSTM) and gated recurrent units

(GRU), gained prominence [20]. These models processed text in sequence, enabling them to capture word order and short-term dependencies—an evident advancement over traditional bag-of-words approaches [21]. However, these architectures struggled with long-range dependencies—difficulty in connecting information across long passages of text—and were often computationally intensive to train on full-length news articles, limiting their effectiveness on complex narratives [22].

2.3. The transformer revolution

A significant breakthrough emerged with the introduction of the transformer architecture and its self-attention mechanism, first proposed by Vaswani *et al.* [23]. The transformer's capability to assess the relevance of all words in a sequence simultaneously—irrespective of their distance—effectively addressed the long-range dependency problem. This led to the development of pre-trained language models like BERT [9], which revolutionized natural language processing (NLP). BERT's pre-training on two unsupervised tasks—the masked language model (MLM) and next sentence prediction (NSP)—allows it to develop a deep, bidirectional understanding of language context. As empirically demonstrated by Ramzan *et al*. [12], this gives it a distinct advantage over both traditional models and unidirectional models like LSTMs, especially when generalizing to new, unseen data where context is paramount.

Following this shift, the dominant paradigm has been the development and end-to-end fine-tuning of increasingly complex BERT-based architectures. Variants like RoBERTa [24], which optimized BERT's training methodology, and ALBERT [25], which introduced parameter-reduction techniques for efficiency, pushed performance even higher. Researchers quickly adapted these models for fake news detection. For instance, Kaliyar *et al*. [26] proposed FakeBERT, combining BERT with a CNN to enhance feature extraction, while Jwa *et al*. [27] introduced exBAKE, which augments BERT's pre-training with a large corpus of news articles to improve domain-specific knowledge. The versatility of BERT is further highlighted by its use as a core textual analysis component in multimodal systems, which analyze the coherence between a news article's text and its accompanying images to detect inconsistencies [28]. In parallel, other research avenues have explored stance detection [29], [30], rumor propagation on social media [31], and advanced architectures like graph neural networks (GNNs) to model relationships within news content [32], [33].

2.4. Hybrid models and interpretability

Despite impressive performance of these end-to-end models, a significant and widely acknowledged drawback is their inherent lack of interpretability. A fine-tuned transformer with millions of parameters functions as a "black box", making its decision-making process opaque to human users [13], [14]. This "interpretability crisis" has become a major focus of the XAI movement in NLP [15]. Recent work by Li *et al.* [34] further emphasizes that while models like BERT are powerful, their black-box nature can hinder trust and adoption in critical journalistic applications.

This challenge has motivated two parallel lines of research. The first focuses on developing post-hoc explainability methods to probe these complex models, using techniques like local interpretable model-agnostic explanations (LIME), which approximates the model's behavior locally, or Shapley additive explanations (SHAP), which applies game-theoretic principles to allocate importance scores to features [18], [19]. For example, Szczepański *et al*. [35] developed a new method specifically to provide explanations for BERT-based fake news classifiers after they have made a prediction.

The second line of research, which this work contributes to, involves designing hybrid systems that are more transparent by design. This approach decouples the feature extraction from the classification stage. The powerful but opaque transformer is used solely to generate high-quality semantic embeddings, which are then fed into an inherently more transparent and efficient classifier. Studies have shown that combining deep learning embeddings with tree-based models like XGBoost or light gradient boosting machine (LightGBM) can yield competitive or even superior performance in various text classification tasks, often with a fraction of the inference time [36]–[38]. This work builds upon this principle of decoupled architectures, applying it to fake news detection to create a system that retains the state-of-the-art performance of modern transformers while providing a clearer, more auditable, and ultimately more trustworthy decision-making process.

3. METHODOLOGY

To conduct a fair and rigorous comparison between a standard end-to-end transformer model and our proposed hybrid architecture, a unified experimental workflow is designed. This process begins with

data preparation and culminates in a comparative evaluation of two distinct modeling architectures built upon the same foundational language model. The architectural overview of our proposed hybrid system is depicted in Figure 1. The diagram illustrates the decoupled two-stage process: first, deep feature extraction using a fine-tuned DistilBERT model to generate semantic embeddings, followed by a transparent classification stage using an XGBoost model.

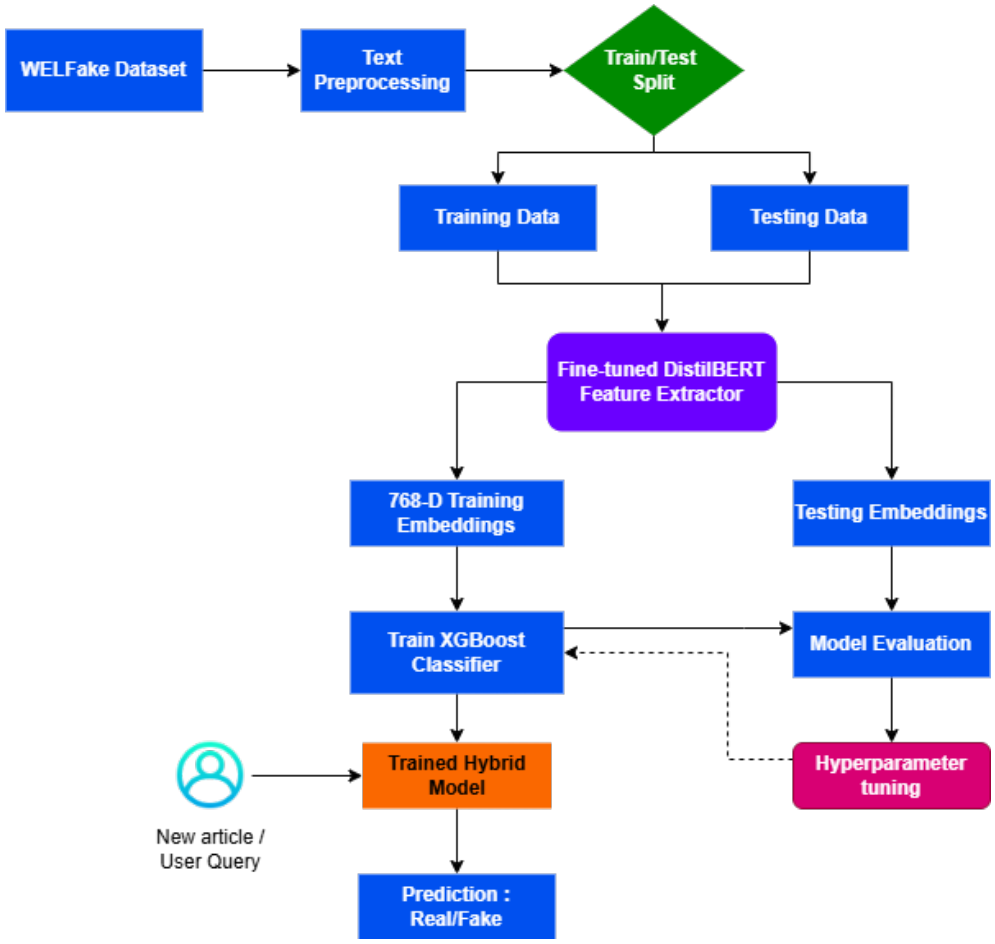

Figure 1. An architectural overview of the proposed hybrid fake news detection system

3.1. Dataset and preprocessing

This study utilizes the complete WELFake dataset [39], a large and balanced corpus for fake news research containing articles from various sources. The dataset's primary features are the raw text of the articles, contained in 'title' and 'text' columns, alongside a binary 'label'. This approach uses only this textual data, foregoing any reliance on handcrafted or metadata features. The initial dataset contains 72,134 news articles. This preprocessing pipeline was designed to ensure data quality and prepare the text for transformer-based analysis. First, all rows containing null values in critical fields such as the title or text were removed, resulting in a cleaned dataset of 71,537 articles, with 34,704 labeled as real and 36,833 as fake. Next, to provide the model with maximum context, the article title and text fields were concatenated into a single full_text input, separated by a special [SEP] token. This allows the model to leverage signals from both the headline and the body of the article. Finally, the full, cleaned dataset was partitioned into a training set (80%) and a final, held-out test set (20%). This resulted in a training corpus of 57,229 articles and a test set of 14,308 articles. Stratification was employed during this split to ensure that the original distribution of real and fake news labels was preserved in both partitions. This test set was kept entirely separate and was only used for the final evaluation of the trained models to guarantee an unbiased assessment.

3.2. Model architectures for comparison

Both architectures are built upon the distilbert-base-uncased pre-trained language model, a lighter and faster version of BERT that retains most of its performance [40]. This choice allows for a direct and fair comparison of the architectural approaches. By using the same distilled model as a foundation, the performance and interpretability differences attributable solely to the architectural choice—end-to-end fine-tuning versus our decoupled hybrid system—can be isolated.

3.2.1. Baseline model: end-to-end fine-tuning

The baseline represents the standard high-performance approach in modern NLP. The DistilBertForSequenceClassification model are utilized from the Hugging Face Transformers library, which appends a classification head (a linear layer) to the core DistilBERT model. This model contains approximately 66 million trainable parameters. This entire model, including the transformer body and the new classification layer, was trained end-to-end on the 57,229-sample training set. Training was conducted for a maximum of 10 epochs, with an early stopping callback monitoring the validation loss on a subset of the training data. This mechanism ensures that selection of the optimal model checkpoint, effectively preventing overfitting.

3.2.2. Proposed model: hybrid architecture

The proposed model is a decoupled, three-stage hybrid architecture designed to achieve high performance while enhancing model transparency and interpretability. The first stage mirrors the baseline's fine-tuning process. The objective here is not to create the final classifier, but to adapt the internal parameters of the DistilBERT model. This is accomplished through the self-attention mechanism, enabling the model to evaluate the relative importance of various words within the input text. The core of this is the scaled dot-product attention, given by (1).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Where Q (query), K (key), and V (value) are matrices derived from the input embeddings, and d_k is the dimension of the keys. This process transforms the model from a general-purpose language model into a domain-specific expert. After fine-tuning, the classification head of the transformer is discarded. The specialized DistilBERT encoder is then used as a powerful feature extractor [41]. All articles are processed through this model, and for each article, the 768-dimensional embedding from the final hidden state of the special `[CLS]` token is extracted. This vector serves as a rich, dense feature representation. Finally, an XGBoost classifier is trained on these embeddings [42]. XGBoost optimizes an objective function that combines a loss term and a regularization term, as defined in (2).

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

Where l is the loss function, and Ω is a regularization term that penalizes model complexity.

3.3. Experimental setup

All experiments were carried out in the Kaggle notebook environment using a Tesla T4 GPU to facilitate model training and acceleration. The implementation was carried out using Python with the PyTorch, transformers, XGBoost, and Scikit-learn libraries. For DistilBERT fine-tuning, the model was trained for a maximum of 10 epochs with a batch size of 16, using the AdamW optimizer and an early stopping callback with a patience of 1 to prevent overfitting. For the XGBoost classifier, which is significantly less complex, a model with n_estimators=200 and a maximum tree depth of 5 was trained, providing a robust configuration. To ensure the reproducibility of the results, a consistent random seed (SEED=42) was used throughout all stages of data partitioning and model training.

3.4. Evaluation framework

The final performance of both models was assessed on the held-out test set. A suite of standard classification metrics are employed: overall accuracy, precision, recall, and F1-score. These are defined as (3) to (5).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Where true positives (TP), false positives (FP), and false negatives (FN), respectively. The confusion matrix and the feature importance scores also analyzed from the XGBoost model to evaluate its interpretability [43].

4. RESULTS AND DISCUSSION

The experimental results, derived from training on the full cleaned WELFake dataset, demonstrate the exceptional performance of the hybrid BERT-XGBoost model. The findings confirm that this architecture achieves state-of-the-art accuracy while offering significant, tangible advantages in model analysis and interpretability. The following sections will detail these findings, beginning with a quantitative performance comparison, followed by an analysis of the training dynamics, and culminating in an in-depth examination of the model's interpretability.

4.1. Quantitative performance comparison

The detailed performance metrics for our proposed hybrid model are presented in Table 1. On the held-out test set of 14,308 articles, the model achieved an outstanding overall accuracy of 99.76%. The precision, recall, and F1-scores are exceptionally high and well-balanced at 0.9975 or higher for both the "Real" and "Fake" classes. This balance is a strong indicator of a robust classifier that is not biased towards one class and performs reliably on both positive and negative samples. The high precision for the "Fake" class, in particular, means that when the model flags an article as fake, its judgment is highly trustworthy, while the high recall demonstrates its effectiveness in capturing the vast majority of misinformation.

Table 1. Detailed performance of the hybrid model on the full test set

Class	Precision	Recall	F1-score	Support
Real (0)	0.9973	0.9977	0.9975	7006
Fake (1)	0.9978	0.9974	0.9976	7302
Accuracy		0.9976		
Macro Avg	0.9975	0.9976	0.9976	14308
Weighted Avg	0.9976	0.9976	0.9976	14308

To contextualize this performance, Table 2 compares our results against several other notable BERT-based approaches from the literature. This model significantly outperforms baseline BERT implementations [12] and demonstrates a notable improvement over other fine-tuned models on different datasets [13]. Most importantly, the hybrid model's accuracy of 99.76% is highly competitive with, and even surpasses, state-of-the-art specialized architectures like FakeBERT [26], which reached 98.90% accuracy. This result is particularly compelling as our model achieves this top-tier performance on the full WELFake dataset—which is significantly larger and more diverse than the corpora used in several baseline studies—while also offering the crucial benefits of transparency. Furthermore, the end-to-end fine-tuned DistilBERT model achieved a validation accuracy of 99.72% on the same data. This confirms that our decoupled hybrid approach sacrifices no discernible predictive power; in fact, the slight performance edge suggests a synergistic benefit, where the specialized XGBoost classifier is able to leverage the rich embeddings more effectively than a standard linear classification head.

Table 2. Comparative analysis of fake news detection models with dataset and model context

Model/Approach	Dataset	Articles	Class balance (R/F)	Model size	Acc. (%)
BERT (Baseline) [12]	Custom small	˜2,000	Approx. Bal.	˜110M	84.00
RoBERTa (Fine-Tuned) [13]	NELA-GT-2022	˜8,000	Unbalanced	˜125M	89.68 (F1)
FakeBERT (BERT+CNN) [26]	Kaggle/ISOT	˜45,000	Bal. (53/47)	$> 110M$	98.90
End-to-End DistilBERT	WELFake	71,537	Bal. (48/52)	˜66M	99.72
Hybrid BERT+XGBoost	WELFake	71,537	Bal. (48/52)	˜66M+XGB	99.76

To provide qualitative evidence of the model's performance, Table 3 showcases the headlines from several articles that were correctly classified by the hybrid system. While the model made its predictions using

the full concatenated title and text, the headlines alone often reveal the stark contrast in linguistic style between the two classes. The fake news examples are characterized by sensationalism, including the use of all-caps (e.g., PRICELESS!), clickbait framing (e.g., WATCH:), and emotionally charged, non-journalistic language. In contrast, the real news headlines, which are traceable to credible outlets, maintain a factual and objective tone. This demonstrates the model's capability to learn and generalize based on these crucial stylistic cues, which are often most prominent in the headline.

Table 3. Sample headlines from correctly classified articles on the test set

Headline	True label	Predicted label
Examples of correctly classified fake news		
PRICELESS! ANTI-TRUMP RIOTER THROWS TANTRUM When Arrested: "I want…I want…I want!" [Video]	Fake	Fake
WATCH: embattled GOP senator just killed his campaign with this racist remark	Fake	Fake
Examples of correctly classified real news		
Hillary Clinton Dines with Her 'SNL' Impersonator Kate McKinnon	Real	Real
Islamist militants kill six soldiers in southern Philippines	Real	Real
Snap shares leap 44% in debut as investors doubt value will vanish	Real	Real

Note: The model's prediction was based on the full article text; headlines are shown here for concise illustration of stylistic differences

4.2. Training dynamics and error analysis

The high performance is underpinned by a robust and efficient fine-tuning process. The validation loss was monitored across epochs, as shown in Figure 2. The loss reached its minimum after the third epoch, and our early stopping callback correctly selected this checkpoint for the final feature extraction model. This rapid convergence demonstrates the power of transfer learning; the pre-trained DistilBERT model required only a few epochs to specialize for the task. The subsequent slight rise in validation loss in the fourth epoch confirms that the early stopping mechanism was crucial in preventing overfitting and selecting the most generalizable model.

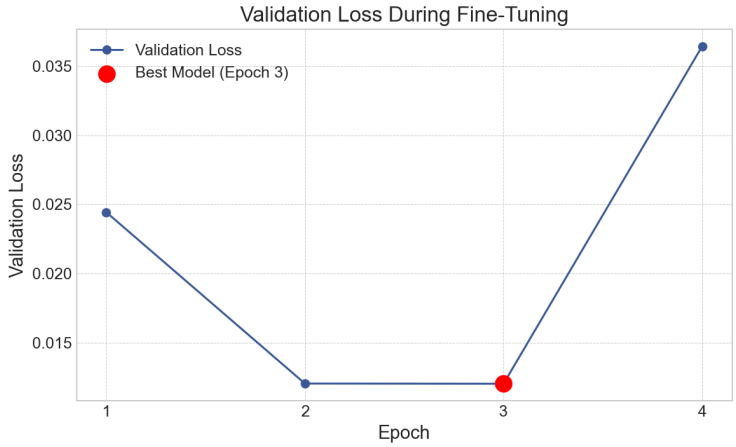

Figure 2. Validation loss per epoch during DistilBERT fine-tuning

A granular analysis of the hybrid model's error profile is provided by the confusion matrix in Figure 3. The model made a total of only 35 misclassifications out of 14,308 test samples. Critically, the number of false negatives—the most dangerous error type where a fake article is misclassified as real—was extremely low at just 19 instances out of over 7,300 fake articles. This demonstrates the model's high sensitivity in catching misinformation. Similarly, the model produced only 16 false positives, where legitimate news is incorrectly flagged as fake. While less harmful, minimizing this error is important for maintaining trust in credible sources. Following this, the classifier's calibration is illustrated in Figure 4, which shows that the optimal F1-score is achieved at a threshold of approximately 0.59. The flatness of the curves across a wide range of thresholds indicates a very robust and well-separated classifier. This signifies that the feature embeddings for

the “Real” and “Fake” classes are distinctly clustered, allowing the XGBoost model to find a clear and stable decision boundary, making its performance not overly sensitive to the exact threshold choice.

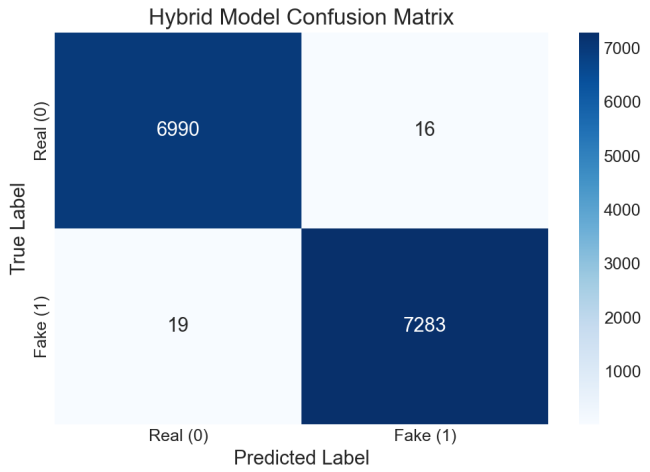

Figure 3. Confusion matrix for the proposed hybrid model on the full test set

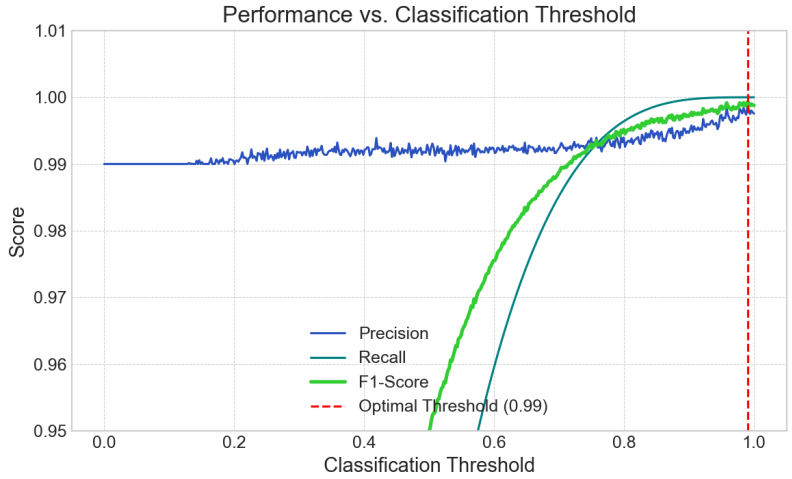

Figure 4. Precision, recall, and F1-score as a function of the classification threshold for the hybrid model

4.3. Model interpretability with Shapley additive explanations

The hybrid architecture’s primary advantage is its inherent interpretability. Unlike “black-box” end-to-end models, the XGBoost classifier allows detailed inspection of features provided by the fine-tuned transformer. Figure 5 presents a SHAP summary plot, providing a far richer view than standard feature importance charts [19]. Each dot in the figure represents a sample from the test set; its color indicates the feature’s value (red is high, blue is low), and its position on the x-axis shows its impact on the prediction score. Crucially, the features shown (e.g., Feature_512) are not pre-defined linguistic inputs, but dimensions within the 768-dimensional embedding space generated by DistilBERT.

The SHAP plot confirms the importance of features like Feature_512 and Feature_15, revealing their directional impact: high values (red dots) push predictions towards “Fake”, while low values (blue dots) favor “Real”. Instance-level explanations, shown in Figure 6, bridge abstract features to concrete analysis. At this figure, features in red (e.g., Feature_512, Feature_15) increase the likelihood of a “Fake” prediction, pushing the output value higher from the base value. Features in blue push it lower. Qualitative inspection of articles

with high positive Feature_512 values reveals common misinformation traits, such as sensationalist headlines, emotionally charged adjectives (e.g., outrageous, shocking), and conspiratorial framing. This suggests the dimension acts as a high-level detector for inflammatory language. While full linguistic mapping remains future work, this demonstrates how our hybrid model enables genuinely explainable fake news detection [43].

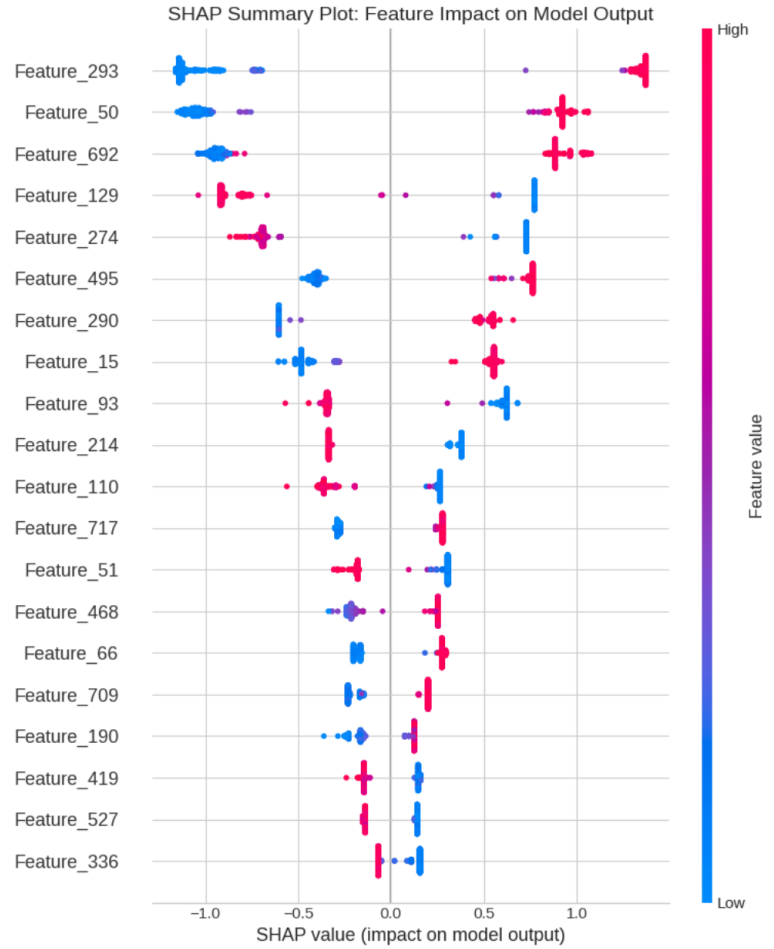

Figure 5. SHAP summary plot showing the impact of the top 20 features on the model's output

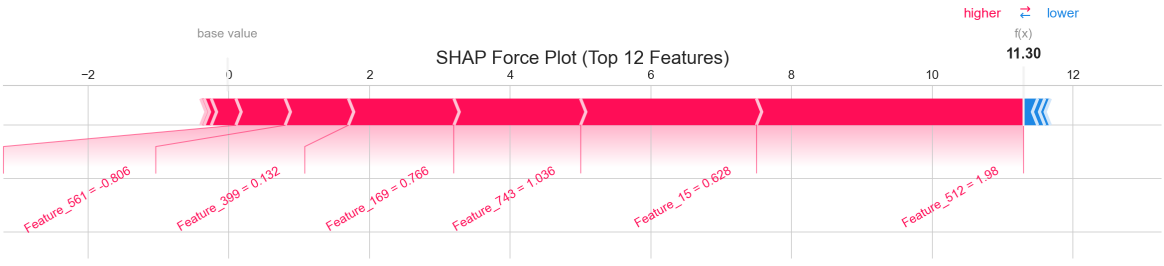

Figure 6. SHAP force plot for a single fake news article

While SHAP identifies prediction drivers, ensuring explanation faithfulness remains a core XAI challenge [43]. Furthermore, despite exceptional accuracy, the model's long-term production utility depends on robustness against adversarial attacks targeting transformer filters [44]. To transition to a comprehensive

trust-verification system, future iterations could integrate multimodal analysis techniques [45], [46], automated claim verification frameworks [47]–[49], and broader rumor detection datasets [50], creating a multi-layered defense against misinformation.

5. CONCLUSION

This paper conducted a comprehensive evaluation of a hybrid BERT-XGBoost model, achieving a state-of-the-art accuracy of 99.76% on the WELFake dataset. The results confirm that decoupling feature extraction from classification provides a vital layer of transparency via SHAP analysis without sacrificing predictive power, addressing the inherent "black-box" limitations of standard transformers. However, this study is limited by its validation on a single dataset and that the model's robustness against sophisticated adversarial attacks remains to be fully investigated. Future work will focus on enhancing this robustness, mapping embedding dimensions to concrete linguistic patterns like sensationalism, and validating this hybrid template in other critical domains such as medical misinformation and hate speech detection. Ultimately, this decoupled architecture establishes a superior and more trustworthy strategy for high-stakes text classification where both accuracy and accountability are paramount.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the Management of Rashtreeya Sikshana Samithi Trust (RSST), as well as the principal and vice principal of RV College of Engineering, Bengaluru, India, for their continuous support and encouragement throughout this research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nishant Vasantkumar Hegde	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Suneesh Bare		✓	✓	✓	✓	✓	✓	✓		✓				
Namruth Reddy				✓	✓					✓		✓	✓	
Rajat Gondkar Aravinda		✓	✓	✓	✓	✓	✓	✓		✓				
Minal Moharir		✓		✓	✓					✓		✓	✓	
Aamir Ibrahim			✓	✓		✓				✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AND CODE AVAILABILITY

The data that support the findings of this study are from the WELFake dataset, which is publicly available with the original study cited as reference [39]. The complete source code for data preprocessing and

model training has been made publicly available in a GitHub repository. The repository can be accessed at https://github.com/kernelops/fakenews-detection-bert-xgboost-hybrid.git.

REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146-1151, Mar. 2018, doi: 10.1126/science.aap9559.

[2] D. M. J. Lazer *et al.*, "The science of fake news," *Science*, vol. 359, pp. 1094-1096, Mar. 2018, doi: 10.1126/science.aao2998.

[3] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156151-156170, 2021, doi: 10.1109/ACCESS.2021.3129329.

[4] X. Zhou and R. Zafarani, "A survey of fake news: fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, Sep. 2021, doi: 10.1145/3395046.

[5] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, Jan. 2018, doi: 10.1002/spy2.9.

[6] D. S. R. Krishna, D. S. V. Vasantha, and K. M. Deep, "Survey on fake news detection using machine learning algorithms," *International Journal of Engineering Research & Technology*, vol. 9, no. 8, Jun. 2021, doi: 10.17577/IJERTCONV9IS08026.

[7] C. C. Aggarwal, "Text classification: basic models," in *Machine Learning for Text*, Cham, Switzerland: Springer, 2018, pp. 113–157, doi: 10.1007/978-3-319-73531-3_5.

[8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: a data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[10] Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, Nov. 2016, doi: 10.1613/jair.4992.

[11] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 15–18, doi: 10.18653/v1/N19-5004.

[12] A. Ramzan, R. H. Ali, N. Ali, and A. Khan, "Enhancing fake news detection using BERT: a comparative analysis of logistic regression, RFC, LSTM and BERT," i*n 2024 International Conference on IT and Industrial Technologies (ICIT)*, Dec. 2024, pp. 1–6, doi: 10.1109/ICIT63607.2024.10859673.

[13] S. Raza, D. P. Patterson, and C. Ding, "Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data," *Knowledge and Information Systems*, vol. 67, no. 4, pp. 3267–3292, Apr. 2025, doi: 10.1007/s10115-024-02321-1.

[14] C. D. Manning, "Computational linguistics and deep learning," *Computational Linguistics*, vol. 41, no. 4, pp. 701–707, Dec. 2015, doi: 10.1162/COLI_a_00239.

[15] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[16] S. Maham, A. Tariq, M. U. G. Khan, F. S. Alamri, A. Rehman, and T. Saba, "ANN: adversarial news net for robust fake news classification," *Scientific Reports*, vol. 14, no. 1, Apr. 2024, doi: 10.1038/s41598-024-56567-4.

[17] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: explainable fake news detection," in *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2019, pp. 395–405, doi: 10.1145/3292500.3330935.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': explaining the predictions of any classifier," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777, doi: 10.5555/3295222.3295230.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[21] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using bi-directional LSTM-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74–82, 2019, doi: 10.1016/j.procs.2020.01.072.

[22] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000, doi: 10.1162/089976600300015015.

[23] A. Vaswani *et al*., "Attention is all you need," *31st Conference on Neural Information Processing Systems*, 2017, pp. 1-11.

[24] Y. Liu *et al*., "RoBERTa: a robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[25] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: a lite BERT for self-supervised learning of language representations," *2020 International Conference on Learning Representations*, 2020, pp. 1-17.

[26] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11765–11788, Mar. 2021, doi: 10.1007/s11042-020-10183-2.

[27] H. Jwa, D. Oh, K. Park, J. Kang, and H. Lim, "exBAKE: Automatic fake news detection model based on bidirectional encoder representations from transformers (BERT)," *Applied Sciences*, vol. 9, no. 19, Sep. 2019, doi: 10.3390/app9194062.

[28] T. Zhang *et al*., "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection," in *2020 International Joint Conference on Neural Networks*, Jul. 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206973.

[29] T. Aljrees *et al*., "Fake news stance detection using selective features and FakeNET," *PLoS ONE*, vol. 18, no. 7, Jul. 2023, doi: 10.1371/journal.pone.0287298.

[30] L. Tian, X. Zhang, Y. Wang, and H. Liu, "Early detection of rumours on Twitter via Stance transfer learning," in *Advances in Information Retrieval*, vol. 12035, 2020, pp. 575–588. doi: 10.1007/978-3-030-45439-5_38.

[31] Q. Li, Q. Zhang, L. Si, and Y. Liu, "Rumor detection on social media: datasets, methods and opportunities," in *2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2019, pp. 66–75, doi: 10.18653/v1/D19-5008.
[32] A. A. Harby and F. Zulkernine, "A comparative analysis of graph neural networks for fake news detection," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference*, 2023, pp. 1215–1222, doi: 10.1109/COMPSAC57700.2023.00184.
[33] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, "User preference-aware fake news detection," in *44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2021, pp. 2051–2055, doi: 10.1145/3404835.3462990.
[34] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualizing and understanding neural models in NLP," in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 681–691, doi: 10.18653/v1/N16-1082.
[35] M. Szczepański, M. Pawlicki, R. Kozik, and M. Choraś, "New explainability method for BERT-based model in fake news detection," *Scientific Reports*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-03100-6.
[36] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *15th Conference of the European Chapter of the Association for Computational Linguistics*, Apr. 2017, pp. 427–431.
[37] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 328–339, doi: 10.18653/v1/P18-1031.
[38] D. Dhiman, V. Asha, M. Devi, S. Shetty, and U. Gurav, "Hybrid sentiment analysis model combining BERT and gradient boosting for enhanced text classification," in *2025 3rd International Conference on Data Science and Information System (ICDSIS)*, May 2025, pp. 1–6, doi: 10.1109/ICDSIS65355.2025.11070894.
[39] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, Aug. 2021, doi: 10.1109/TCSS.2021.3068519.
[40] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[41] R. Corizzo, E. Zdravevski, M. Russell, A. Vagliano, and N. Japkowicz, "Feature extraction based on word embedding models for intrusion detection in network traffic," *Journal of Surveillance, Security and Safety*, vol. 1, pp. 140-150, 2020, doi: 10.20517/jsss.2020.15.
[42] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[43] R. K. Sinha, "Book review: Christoph Molnar. 2020. Interpretable machine learning: a guide for making black box models explainable," *Metamorphosis: A Journal of Management Research*, vol. 23, no. 1, pp. 92–93, Jun. 2024, doi: 10.1177/09726225241252009.
[44] A. Luz and E. Frank, "Adversarial attacks on BERT-based fake news detection models," *EasyChair Preprint*, 2024.
[45] X. Shen, M. Huang, Z. Hu, S. Cai, and T. Zhou, "Multimodal fake news detection with contrastive learning and optimal transport," *Frontiers in Computer Science*, vol. 6, Nov. 2024, doi: 10.3389/fcomp.2024.1473457.
[46] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection," *LREC 2020 - 12th International Conference on Language Resources and Evaluation*, pp. 6149–6157, Mar. 2020.
[47] H. Liu *et al*., "Retrieval augmented scientific claim verification," *JAMIA Open*, vol. 7, no. 1, 2024, doi: 10.1093/jamiaopen/ooae021.
[48] J. Chen, G. Kim, A. Sriram, G. Durrett, and E. Choi, "Complex claim verification with evidence retrieved in the wild," in *2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024, pp. 3569–3587, doi : 10.18653/v1/2024.naacl-long.196.
[49] N. Jafari and J. Allan, "Robust claim verification through fact detection," Jul. 2024, *arXiv: 2407.18367*.
[50] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PLoS ONE*, vol. 11, no. 3, Mar. 2016, doi: 10.1371/journal.pone.0150989.

BIOGRAPHIES OF AUTHORS

Nishant Vasantkumar Hegde is an undergraduate student in the Department of Computer Science and Engineering at RV College of Engineering, Bengaluru, India. He has published research papers in IEEE journals and international conferences, including a novel encryption algorithm presented at IIT Indore, and has authored journal articles in IEEE Access. He interned at Samsung R&D Institute under the Samsung PRISM program, focusing on image processing. His research interests include artificial intelligence, machine learning, cybersecurity, and cryptography. He can be contacted at email: hegde.nishant2005@gmail.com.

Suneesh Bare is an undergraduate student in the Department of Computer Science and Engineering at RV College of Engineering, Bengaluru, India, specializing in cyber security. He has a keen interest in embedded C, machine learning, and data structures. He is continuously exploring ways to apply AI in real-world security problem-solving. He can be contacted at email: suneeshbare.cy23@rvce.edu.in.

Namruth Reddy is a Senior Security Engineer at NVIDIA Corporation, Santa Clara, United States. He holds expertise in cybersecurity and large-scale system architecture. In this work, he provided crucial industry perspective and guidance, helping to shape the research direction and ensure the practical relevance of the proposed hybrid model. He can be contacted at email: reddynamruth@gmail.com.

Rajat Gondkar Aravinda is an undergraduate student in the Department of Computer Science and Engineering at RV College of Engineering, Bengaluru, India. His research interests include blockchain systems and decentralized applications, artificial intelligence and machine learning, and scalable cloud-based systems. He has worked on applied AI systems for real-time surveillance and intelligent assessment platforms, full-stack architectures for deploying ML models in production environments, as well as blockchain-enabled platforms for secure credentials and creator monetization. He can be contacted at email: rajatga.cs23@rvce.edu.in.

Minal Moharir is currently a professor with the Department of Computer Science and Engineering , R. V. College of Engineering, Bengaluru, India, have teaching experience of 23 years, her specialization includes computer networks and security, machine/deep learning and image processing, and parallel computing. She has published more than 60 papers in reputed journals/conferences. She has also executed sponsored projects worth INR 70 lakhs funded from various agencies nationally and internationally. She is a member of IEEE and ACM. She can be contacted at email: minalmoharir@rvce.edu.in.

Aamir Ibrahim is an undergraduate student in the Department of Computer Science and Engineering at RV College of Engineering, Bengaluru, India. He has co-authored a research paper on novel encryption for resource-constrained systems. His interests include cybersecurity, AI, and FinTech. He has worked on projects involving AI heads-up displays and ECG processing. He can be contacted at email: aamiribrahim.cy23@rvce.edu.in.