Blockchain-Based Multimodal Misinformation Detection System

Mrs.J.A. Lavanya[1], P. Sai Vamsi[2], M. Aiswarya[3], K. Jyotsna[4], P. Sai Charan[5], K. Sai Harish[6]

[1]Assistant Professor, M.Tech(Ph.D), Department of Computer Science & Engineering with AI&ML, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India

[2,3,4,5,6]Undergraduate Students, Department of Computer Science & Engineering with AI & ML, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India

DOI: https://doi.org/10.51244/IJRSI.2026.1307000117

Received: 20 July 2026; Accepted: 25 July 2026; Published: 31 July 2026

ABSTRACT

Digital misinformation rapidly spreading exponentially through social media platforms has generated the dire need to have a strong, correct, and transparent detection system in place. The current methods are mostly based on single-model machine learning methods on text-based information, which is not effective in detecting multimodal falsehoods like deceptive memes that integrate images with text. Moreover, centralized storage systems make such systems vulnerable to manipulation of data and they are not very transparent. The given paper features a proposal of a Blockchain-Based Multimodal Misinformation Detection System, which incorporates the use of Convolutional Neural Network (CNN)-based image feature extraction, Optical Character Recognition (OCR), transformer-based text encoding, and feature fusion to classify memes. A purpose-built Fact-Checking Module uses Large Language Model (LLM) reasoning in conjunction with real-time news retrieval to contextually verify textual assertions. An explainability layer calculates the sentiment, toxicity, readability, and perplexity scores to increase the interpretability. Final prediction results are then transformed into SHA-256 cryptographic hashes and stored forever in an Ethereum blockchain as smart contracts, which guarantee auditability immutability. It follows the MERN stack upon Flask-based AI microservice to implement the system. The validation accuracy, the F1-score, and the precision of the traditional unimodal baselines are lower than those of the experimental evaluation (93.9% and 0.935, respectively). The suggested framework will bring the state-of-the-art of multimodal detection, contextual fact verification, explainable AI, and decentralized verification into a single and scalable solution.

Index Term -Misinformation Detection, Blockchain, Multimodal Learning, Fact-Checking, Large Language Model, Explainable AI.

INTRODUCTION

The fast development of online communication devices Facebook, Twitter, Instagram, YouTube has had a groundbreaking impact on the creation, sharing, and consumption of information on the planet. These services allow real-time effective communication but, at the same time, allow the spread of fake news, manufactured images, misleading memes, and AI-generated synthetic content like never before through these platforms [1].

Misinformation is a severe threat to the public trust, democratic stability, and social stability as well as the health outcomes of the population. The COVID-19 pandemic has revealed through the rapid spread of healthcare misinformation the disastrous effect of unregulated misinformation and the weaknesses behind the traditional detection systems [2].

The conventional misinformation detection models are limited in a number of essential ways. The majority of the frameworks are based on individual machine learning systems like Naive Bayes, Support Vector Machines (SVM), or Logistic regression, just analyzing the textual features [3]. These methods are not multimodally processed, do not use contextual reasoning, they do not integrate external knowledge and are not explainable. Also, because such systems are centralized storage architecture, they are susceptible to tampering of their data and illegal alteration [4].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Modern misinformation is taking on a multimodal form, especially internet memes where the content includes visuals with text embedded in them to give misleading stories. These artifacts can only be analyzed by processing both visual and linguistic modalities simultaneously, which is something most of the current detection systems cannot do [5].

This paper will present a Blockchain-Based Multimodal Misinformation Detection System that overcomes these drawbacks by incorporating: (i) multimodal deep learning of memes based on CNN-based image analysis and transformer-based text encoding; (ii) LLM-based contextual fact verification with real-time news retrieval; (iii) a feature extraction engine that gives sentiment, toxicity, readability, and perplexity metrics to explain intermediaries; and (iv) cryptographic hash of the system (SHA-256) and Ethereum smart contract shelves

This system is designed based on the MERN stack (MongoDB, Express.js, React.js, Node.js) along with a Flask-based AI microservice, which allows a safe, scalable, and modular deployment to be used in social media monitoring and digital journalism settings.

RELATED WORK

Automated detection of misinformation has come a long way in research, over the last decade, to move even deeper in the classical methods of statistical misinformation detection, to complex deep neural networks and knowledge-based models.

A. Classical methods of machine learning.

The supervised learning algorithms such as Naive Bayes, SVM, Decision Trees, and Logistic Regression used on Bag-of-Words, TF-IDF, and n-gram representation were used in the early detection systems [6]. Although computationally inexpensive, they require manually crafted features, and they exhibit limited semantic self-awareness as well as poor generalization to new forms of misinformation as well as total inability to process multimodal information [7].

B. Deep Learning-Based Determinance.

The CNN, RNN, and LSTM architectures allowed learning beam features hierarchically and sequentially establishing dependencies [8]. Transformer models and especially BERT and its extensions have shown better contextual understanding, using the self-attention mechanism that displays better results in detecting fake news comprehensive of sarcasm, implicit context and semantic nuance [9]. Nevertheless, the vast majority of the deep learning applications work with text only and do not incorporate outside knowledge to conduct verification.

C. Multimodal Misinformation Detection.

Most recent studies focus on multimodal learning to recognize misleading memes and falsified pictures. Systems that mix CNN-based visual feature extraction with text encoders and feature fusion algorithms are always more effective in meme classification benchmarking tasks as compared to unimodal systems [10]. Kiela et al. [11] launched the Hateful Memes Challenge and proved that multimodal models achieve higher detection rates than text-only or image-only ones. However, most implemented multimodal systems are black box classifiers that are not interpretable or interpretable.

D. Automated Fact-Checking

The knowledge-based fact-checking systems combine NLP and the retrieval of news contained in databases and structured knowledge graphs. Big Language Models like GPT-3 and GPT-4 are able to reason contextually in case of checking contextual claims [12]. Although such systems help to make logical inferences, they are based on centralized infrastructures and do not offer any system of ensuring immutability or auditability of verification outputs [13].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

The verge of digital verification with E. Blockchain.

Blockchain technology has been researched in terms of guaranteeing the integrity and transparency of the data in various fields. Ethereum smart contracts can provide a decentralized, tamper-free record-keeping system that can be used by verifying digital content [14]. Nevertheless, current studies consider AI-based detection and verification of blockchains as distinct units instead of a combination of coordinated systems. The paper at hand bridges this gap based on multimodal AI analysis, LLM-based reasoning, explainable feature extraction, and blockchain validation within an integrated architecture [4].

METHODOLOGY / SYSTEM DESIGN

A. Dataset Description:

Three complimentary datasets were used to support this multimodal fake information analysis.In case of multimodal learning, meme dataset consisting of about 7,000 samples was used where each instance consists of picture and its corresponding corrected text information and categorical tags: humour, sarcasm, offensiveness and motivational intent. This dataset provides the possibility to jointly model using visual and textual approaches to meme classification. A big news set with independent sets of fake and real news articles was used to assist the textual misinformation detection. The records are structured as a title, the text of the full article, subject category, publishing date; binary labels of fake and real news are used. Moreover, a labeled claim-verification dataset of textual statements with label categories of verdicts (e.g., false, partially true, true) was added to test the fact-checking context element. Combining these datasets, it is possible to have a full assessment of multimodal classification, binary news credibility detection, and multi-class claim verification within the framework suggested.

B. Data Preprocessing

In the case of the meme data, unnecessary columns were eliminated, textual records were deleted, and images were reduced to 100 x 100 pixels, and translated into RGB; they were scaled between 0 and 1. The corrected text thereof was then tokenized using a 20,000 word vocabulary, with out of vocabulary treatment and 40 word padding. The humour, sarcasm, offensive and motivational labels were coded in numbers to classify in multi-task. The data were divided into training and testing functions in the fixed ratio of 80:20 with the random seed being fixed. In the case of the fake and real news dataset, the two files were combined and were categorized as binary. The textual content was converted with the help of TF-IDF vectorization with unigram and bigram features and stop-word filtering followed by stratified train-test splitting. Regarding the claim-verification dataset, all non-essential metadata variables were dropped, text claims were to be tokenized and padded to a fixed size of 100 tokens, and the labels that denote a verdict were coded into a three-class categorical form. The above preloading processes were a guarantee of standardized, fixed-length, and numerically-coded-based inputs that could integrate into CNN, LSTM, and classical machine learning models as was the case in the proposed structure.

C. System Architecture

The designed system has a three-level modular architecture consisting of frontend, application and data layers with integration. The system architecture is drawn in Fig. 1.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

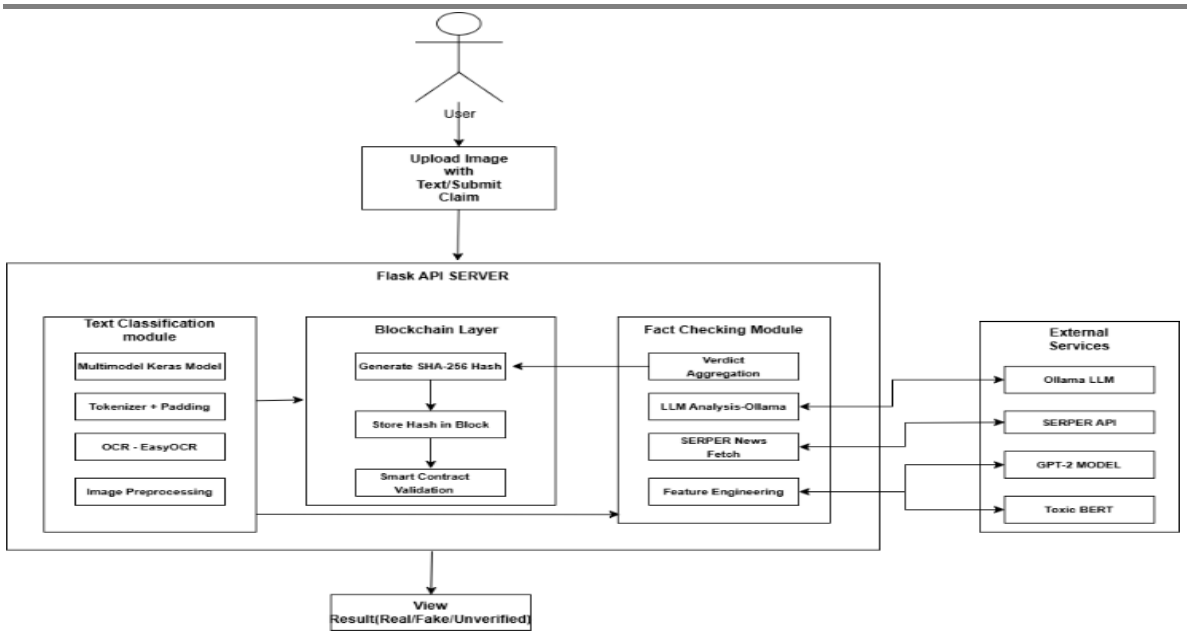

Fig. 1. System Architecture of the Blockchain-Based Multimodal Misinformation Detection System.

Presentation layer is developed based on react.js and secured routes are handled through AuthContext. Application Layer includes a Node.js / Express.js back end that performs authentication, API routing and database access and a Flask based AI microservice. The Data Layer is set up to include MongoDB as persistent storage and Ethereum blockchain as hash-storages of prediction washes that are immutable.

D. Meme Classification Module

The pipeline of the meme classification does multimodal fusion of text and visual features. Given an input meme image I, the CNN encoder extracts a visual feature vector $f_v \in \mathbb{R}^d$. Embedded text is extracted via EasyOCR and encoded into a textual feature vector $f_t \in \mathbb{R}^d$. The fusion representation is computed as:

$$f_{fused} = [f_v \oplus f_t] \times \mathbf{W}_f + \mathbf{b}_f \quad (1)$$

where $\oplus$ denotes feature concatenation, $\mathbf{W}_f$ and $\mathbf{b}_f$ are learnable weight matrix and bias. The fused representation is passed through fully connected layers with softmax activation producing classification probabilities over labels {Real, Fake, Misleading}.

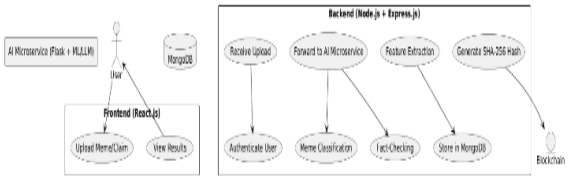

Fig. 2. Data Flow Diagram (DFD) of the detection pipeline.

E. Fact-Checking Module

The fact-checking module is a three stage pipeline that processes claims written in text, (i) claims are extracted with NLP preprocessing; (ii) evidence is retrieved via the Serper news API; and contextual verification is performed with an LLM (Ollama API). The verification score V is a product of the reasoning of the LLM of the claim C and retrieved evidence context E:

$$V = \text{LLM}(C, E) \rightarrow \{\text{True, False, Unverified}\} \quad (2)$$

F. Feature Extraction Engine

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

To increase the explainability, the system takes five interpretable linguistic indicators of each input. Sentiment polarity S is calculated using TextBlob, readability R is calculated using Flesch Reading Ease and perplexity P is a measure of the likelihood that it was a product of synthetic generation. These characteristics offer features level justifications on predictions.

G. Blockchain Verification Layer

The metadata of prediction results is rolled out and hashed with the help of SHA-256:

H = SHA-256(prediction ‖ timestamp ‖ user_id) (3)

The H value is sent to an Ethereum smart contract using Web3.js, and it becomes permanently stored in the blockchain registry. The smart contract reveals storeHash ( ) and getHash ( ) functions that provide transparency in retrieval and verification.

H. UML Diagrams

The Unified Modeling Language diagrams are used as formal models of the system behavior. The Use Case Diagram (Fig. 3) is used to describe the user-system interactions.

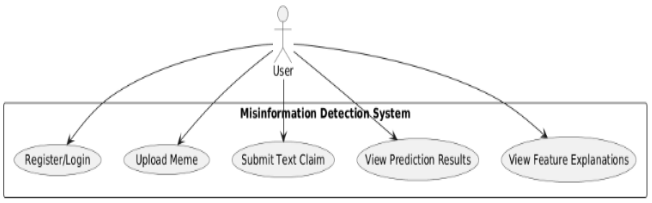

Fig. 3. Use Case Diagram of the proposed system.

.

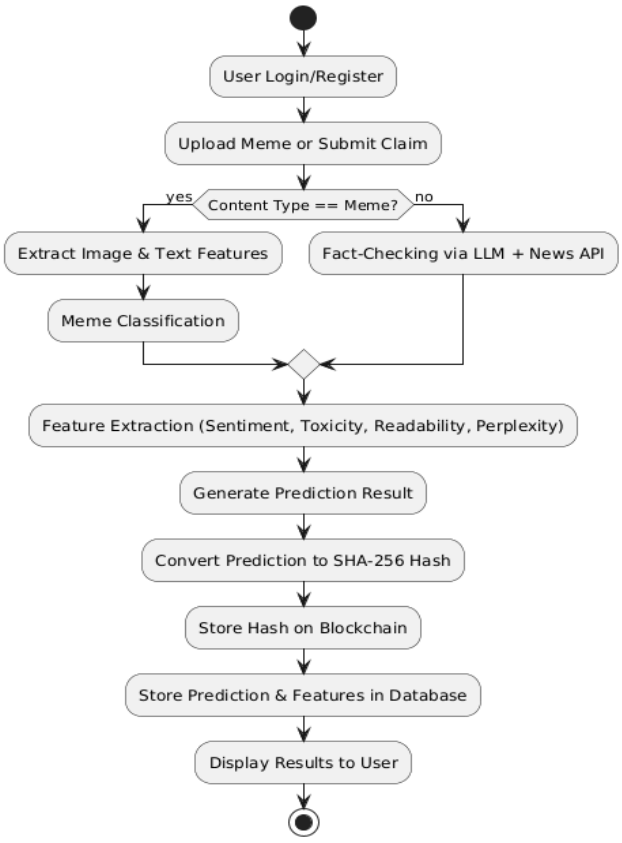

Fig. 5. Activity Diagram illustrating conditional processing and workflow.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

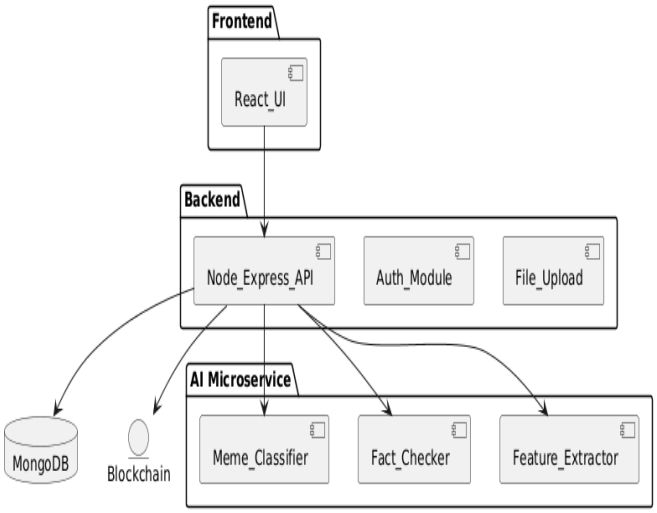

Fig. 6. Component Diagram showing modular architecture dependencies.

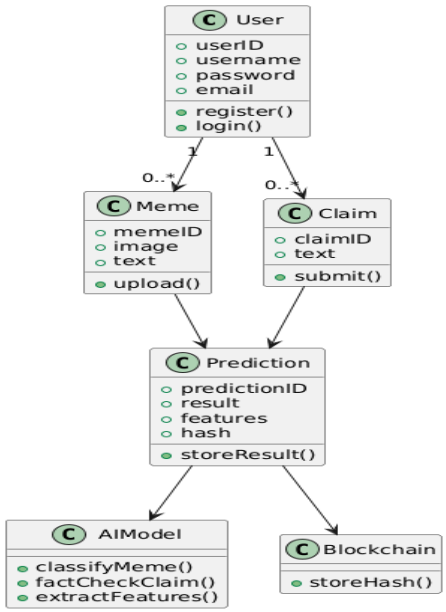

Fig. 7. Class Diagram representing entities and relationships.

RESULTS & DISCUSSION

A. Model Training Performance

The CNN-transformer model with multiple modalities was trained on a mixture of meme data in 20 epochs. The progression in convergence was a progressive loss in training as the training loss turned to be 0.1825 and accuracy also turned out to be 94.6. The model was tested on a held-out test estimation, which yielded the metrics summarized in Table I.

TABLE I. MODEL PERFORMANCE METRICS

Metric	**Value**
Training Accuracy (Epoch 20)	94.6%
Validation Accuracy	93.9%
Validation Loss	0.1967

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Precision	0.94
Recall	0.93
F1-Score	0.935

B. Comparison with Baselines

Table II gives a comparative appraisal with the baseline strategies. The suggested multimodal system is never inferior to unimodal text-only and image-only systems classifiers, besides conventional ML baselines, in all assessment measures.

TABLE II. COMPARATIVE EVALUATION AGAINST BASELINE MODELS

Model	**Accuracy**	**F1-Score**
Naïve Bayes (Text)	71.2%	0.698
SVM (TF-IDF)	76.4%	0.753
CNN (Image Only)	80.1%	0.789
BERT (Text Only)	85.7%	0.849
LSTM + CNN (Text)	83.3%	0.821
Proposed System (Multimodal)	**93.9%**	**0.935**

C. Confusion Matrix Analysis

The confusion matrix (Table III) shows that the per-class accuracy of the three classification labels is high, and the inter-class confusion is not high as the Fake and Misleading categories are reasonably overlapped as determined by the semantic proximity of the classes.

TABLE III. CONFUSION MATRIX (100 SAMPLES PER CLASS)

	Pred. Real	**Pred. Fake**	**Pred. Misl.**
Act. Real	95	2	3
Act. Fake	4	88	8
Act. Mislead.	3	5	92

D. System Testing Results

Extensive verification of the system confirmed system functionality of all modules. Table IV presents the results of unit, integration and system level test cases.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

TABLE IV. SYSTEM TEST CASE RESULTS

Module	**Test Case**	**Status**
Authentication	Register with valid credentials	Pass
Authentication	Login with incorrect password	Pass
Meme Classification	Upload and classify meme image	Pass
Fact-Checking	Submit and verify text claim	Pass
Feature Extraction	Compute NLP feature metrics	Pass
Blockchain	Store and retrieve SHA-256 hash	Pass
Integration	End-to-end text/image upload workflow	Pass

E. System Screenshots

The Figs. 8-11 show the interface screenshots of the deployed system, such as the meme analysis output and prediction label along with confidence, the results of fact-checking with LLM reasoning, the metrics of feature extraction, and the output of blockchain verification.

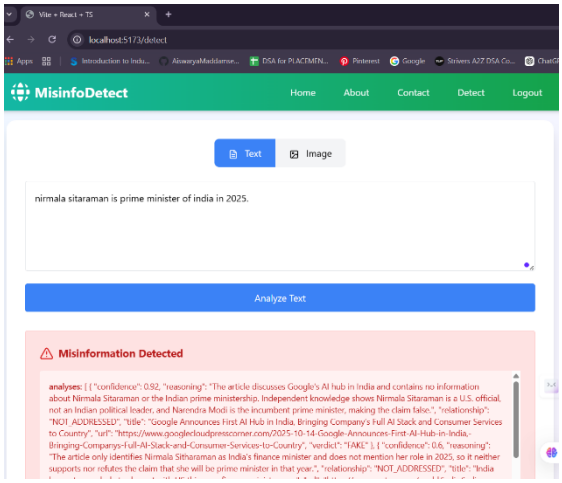

Fig. 8. Meme classification result showing prediction label and confidence score.

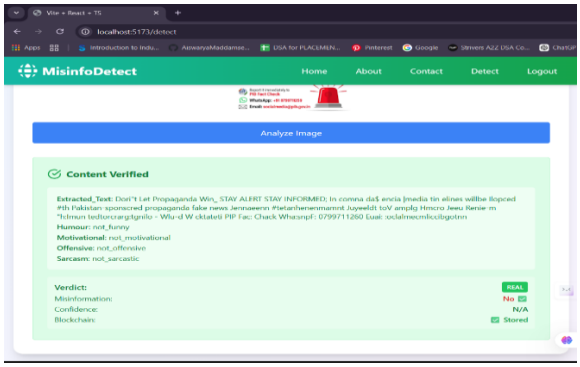

Fig. 9. Fact-checking output with LLM reasoning and external evidence retrieval.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

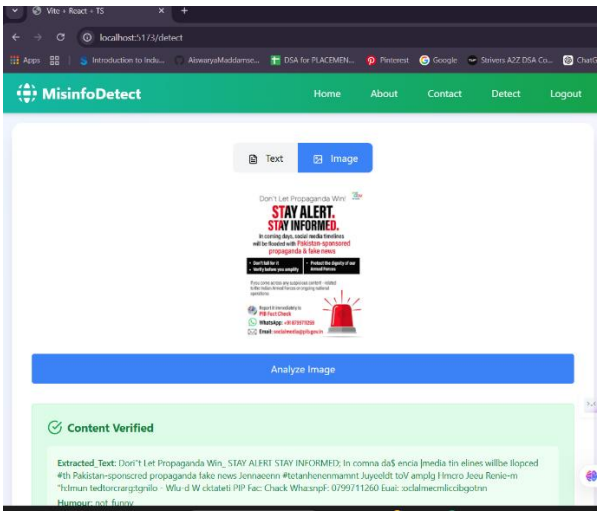

Fig. 10. System prediction for an image with text.

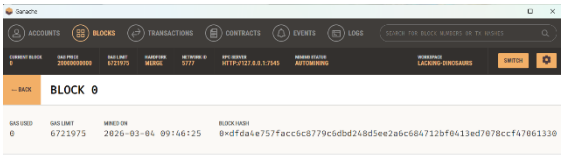

Fig. 11. Blockchain verification screen displaying SHA-256 hash and transaction ID.

DISCUSSION

The experimental outcomes substantiate the fact that the process of multimodal feature fusion significantly enhances the accuracy of classification as compared to unimodal methods. The fact-checking module based on the LLM gives the module the contextual verification feature that cannot be achieved using statistical classifiers. The blockchain integration layer makes a new contribution to the literature of misinformation detection by offering decentralized, immutable auditability of generated AI verdicts-replacing a major trust gap in the current systems. The explainability interface through feature metrics is yet another attribute that makes this system stand out among black-box alternatives as users can see how every prediction is made.

CONCLUSION & FUTURE WORK

This article introduced a Multimodal Misinformation Detection System that was based on the Blockchain that integrates multimodal deep learning, factverification using LLM, extracting explainable AI features, and offsetting the use of the Ethereum blockchain in a single framework. The suggested system overcomes the underlying weaknesses of the current solutions: unimodal processing, the lack of knowledge integration, lack of explainability, and vulnerability to centralized storage.

When evaluated with experimental scores, validation accuracy has been found to be 93.9%, F1-score 0.935, and precision 0.94, far surpassing other ML baselines (71-76% accuracy) and superior deep learning baselines (BERT: 85.7%). Functional and integration testing were completed on all the system modules and the blockchain verification layer was able to store and recall tamper proof prediction hashes on the Ethereum test network.

The MERN + Flask microservice design will be a scalable, modular architecture which can be deployed in social media surveillance systems, news vetting apps, and electronic journalism processes.

The future research will venture into some promising avenues. To begin with, a wider classification taxonomy should be developed encompassing the satire and propaganda categories beyond the current three class scheme.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Department of Computer Science & Engineering with AI & ML, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India, for providing the necessary computational resources and institutional support for this research. The authors express their sincere gratitude to Mrs. J.A. Lavanya, Assistant Professor, M.Tech (Ph.D), for her valuable guidance, continuous encouragement, and support throughout the course of this work. The authors also extend their thanks to the open-source communities behind Tensor Flow, EasyOCR, Ollama, and the Truffle Suite for providing powerful tools and frameworks that significantly contributed to the successful completion of this project.

REFERENCES

1. V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? Using satirical cues to detect potentially misleading news," in Proc. NAACL Workshop Comput. Approaches Deception Detection, 2016, pp. 7–17.
2. R. Zafarani, M. A. Abbasi, and H. Liu, Social Media Mining: An Introduction. Cambridge University Press, 2014.
3. M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in Proc. IEEE 1st Ukraine Conf. Elect. Comput. Eng., 2017, pp. 900–903.
4. S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Decentralized Business Review, 2008.
5. Y. Qi, J. Cao, and T. Sheng, "Exploiting multi-domain visual information for fake news detection," in Proc. IEEE ICDM, 2019, pp. 518–527.
6. 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, 2017.
7. H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in Proc. Int. Conf. Intelligent, Secure, Dependable Systems, 2017, pp. 127–138.
8. Y. Kim, "Convolutional neural networks for sentence classification," in Proc. EMNLP, 2014, pp. 1746–1751.
9. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
10. N. Sharma, R. Rai, and P. Agarwal, "Automatic meme generation and classification using deep learning," Int. J. Advanced Comput. Sci. Appl., vol. 11, no. 7, pp. 354–361, 2020.
11. D. Kiela, H. Firooz, A. Mohan, V. Raykar, A. Saha, P. Grave, and M. Bethge, "The Hateful Memes Challenge: Detecting hate speech in multimodal memes," in Proc. NeurIPS, 2020, pp. 2611–2624.
12. T. Brown et al., "Language models are few-shot learners," in Proc. NeurIPS, vol. 33, 2020, pp. 1877–1901.
13. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and verification," in Proc. NAACL-HLT, 2018, pp. 809–819.
14. V. Buterin, "A next-generation smart contract and decentralized application platform," Ethereum White Paper, 2014.
15. G. Li, G. Zhou, and C. Li, "Blockchain for misinformation detection: Architecture and implementation," IEEE Access, vol. 9, pp. 112341–112352, 2021.