Multimodal DeepFake Detection Through Quantum-Inspired Cross-Language Feature Integration

Dr. A. Naresh[1],R. Sindhu[2],V. Varalakshmi[3],P. Sirisha[4],S. Nandini[5]
Department of CSE, Vignan's Nirula Institute of Technology and Science for women
Palakaluru, Guntur, 522009, Andhra Pradesh, India.

Abstract

The rapid expansion of artificial intelligence has resulted in an increase in AI-generated and deepfake material, posing significant problems to medium legitimacy, security and public trust. To solve this issue, in this offer a hybrid multimodal quantum computing model that combines image, video, audio and text modalities to reliably detect modified content. The system uses specialized feature extraction techniques such as CNNs for visual patterns LSTMs for temporary dynamics, Capsule Networks for structural representation, and BERT for semantic text analysis. The extracted features are combined into a uniform representation and classified using a fully linked decision-making network. Preprocessing is used to maintain uniformity between modalities and prevent false alarm, hence increasing the system's robustness. Quantum-inspired assessment measures are also used to provide more detailed information on detection performance. Suggested paradigm shows great promise for real-time deployment, providing an effective strategy for fighting synthetic media across digital platforms.

Keywords: Multimodal quantum computing, CNN, LSTM, Capsule Networks, BERT.

1.Introduction:

Ensuring the validity of digital material has become more difficult than ever before due to the quick development of synthetic media generation, especially through deepfake technology [1] [2]. Artificially created or altered audio, visual, and textual data, [3] known as "deepfakes" can remarkably resemble genuine people and events. Their growing sophistication has turned them into a weapon for social engineering, political manipulation [4], cybercrime, and disinformation in addition to being a source of amusement [5]. The capacity to consistently identify deepfakes across several modalities has thus emerged as a crucial research topic [6].

Conventional detection systems have mostly depended on artificial intelligence techniques, like transformer-based models for textual discrepancies, recurrent neural networks for sequential data, and convolutional neural networks for visual analysis [7]. Despite their proven efficacy, these strategies are becoming more and more limited [8]. First, complicated high-dimensional correlations across multi modal [9] inputs are difficult for standard deep learning models to capture, particularly when data changes are subtle and adversarially optimized [10]. Second, these models require a lot of resources, including large datasets, a lot of processing power, and lengthy training periods. Third, detection methods frequently fall behind the ever-increasing sophistication of forging tactics due to the rapid advancement of diffusion-based models and generative adversarial networks (GANs) [11].

In this regard, quantum computing emerges as a paradigm-shifting technology that opens up new possibilities for multimodal data analysis and representation [12]. When compared to classical systems, complex correlations can be evaluated with exponential efficiency with quantum mechanics' ability to encode high-dimensional feature spaces into compact quantum states [13]. Quantum-inspired techniques use superposition,[14] entanglement, and quantum interference principles to capture complex interactions across data modalities, in contrast to traditional deep learning techniques [15] that only use linear algebraic

operations over tensors. Following the extraction of modality-specific embeddings, the framework presents a fusion mechanism inspired by quantum mechanics [16]. The suggested approach encodes characteristics into quantum states, where high-dimensional correlations are represented by superposition, as opposed to depending on straightforward concatenation or attention-based fusion [17]. This makes it possible for the system to effectively investigate intricate interdependencies between modalities, including the alignment of spoken content with textual transcripts or the synchronization of lip movements with audio inputs [18]. The fusion process increases sensitivity to tiny multimodal discrepancies that classical systems would miss by utilizing correlations induced by quantum entanglement [19].

Mel-Frequency Cepstral Coefficients (MFCCs),[7] which extract perceptual qualities closely linked with the human auditory system, are used to assess audio streams initially [20]. The temporal dynamics and irregularities in speech patterns are then captured by modeling these aspects with recurrent structures like LSTMs. A CNN-CapsuleNet [21] process is used to evaluate visual input. Convolutional layers identify local patterns, and Capsule Networks maintain spatial hierarchies, allowing the system to identify pixel-level discrepancies and subtle facial micro-expression forgeries. BERT [22] embeddings are used to encapsulate textual content, which frequently contains semantic manipulation in disinformation campaigns, while maintaining linguistic structure and contextual dependencies [23].

The final binary classification which separates real information from altered deepfakes is produced by processing the integrated embeddings through fully linked layers after the fusion stage [24]. By combining the representational power of quantum-inspired computation with the advantages of classical feature extraction, this quantum-hybrid method produces a detection system that is scalable, interpretable, and prepared for the future [25].

This framework's importance stems from both its computational philosophy and multimodal reach. The system leverages quantum probabilistic modeling for decision-making, going beyond brute-force pattern recognition, by integrating quantum principles into the detection process [26]. This change offers resilience against adversarially optimized fake content, permits more compact models, and lowers the danger of overfitting [27]. Moreover, the incorporation of really quantum-enhanced modules [6] may further speed up inference and detection accuracy as quantum technology develops, enabling the system to be tailored to practical uses like cybersecurity, digital forensics, and social media monitoring [28].

In the face of these obstacles, quantum computing shows itself as a revolutionary approach that can solve the fundamental flaws in traditional detection methods. Quantum computing encodes and manipulates data according to the laws of quantum mechanics, as opposed to conventional computational methods that work with binary states [29]. With significant ramifications for high-dimensional, multimodal data analysis,[12] this change introduces essentially new computing possibilities. Massive parallelism is supported at an intrinsic level by quantum superposition, which enables a quantum bit, or qubit, to exist in combinations of several states simultaneously [30]. Instead of processing multimodal information sequentially as in classical systems, [31] this property can be used to simultaneously examine numerous correlations among them for detection tasks. In a similar vein, non-trivial correlations that defy classical probability are established via the quantum entanglement principle, allowing for the representation of much more complex cross-dependencies in data [32].

For instance, in ways that classical models find difficult, entanglement-inspired operations are able to capture subtle alignments between modalities such as audio phonemes and lip articulation in films [33]. Additionally, by suppressing noise and allowing probabilistic amplification of correct decision paths [34], quantum interference provides a way to distinguish subtle differences between manipulated and real signals, differences that would otherwise be obscured by the noise of complex data distributions. Beyond these guidelines, the exponential representational capacity of quantum state spaces simplifies the computational bottlenecks that classical systems encounter when analyzing complex, multimodal deepfake data by enabling the natural encoding of extremely high-dimensional feature vectors into more manageable mathematical structures [35].

After being represented in this quantum-enhanced environment, the fused embeddings are ultimately translated into decision outputs via dense layers, producing a binary classification: real or fake. Crucially, quantum probabilistic modaling is integrated into the decision process, enabling the system to operate more like a probabilistic inference engine and less like a deterministic classifier [36]. Because attackers can no longer take advantage of perfectly deterministic decision boundaries, this adds resilience against adversarial optimization tactics. Therefore, the hybrid design capitalizes on the complementing characteristics of both paradigms: the strong representational and inferential capabilities of quantum encoding coupled with the proven empirical efficiency of classical feature extraction. However, large-scale quantum neural networks are not yet fully realized on existing hardware [37].

The proposed framework ultimately goes beyond addressing a technological issue; it embodies a significant epistemological shift in the approach to deepfakes [38]. It conceptualizes reality verification as a multimodal phenomenon that can be represented in a computational model capable of reasoning about probability and coherence interdependencies at scales that classical computing cannot handle [39]. In this way, it serves as a proactive tool against current generation deepfake threats and also anticipates the future trajectory of adversarial synthetic media [14] in the coming years. As post-quantum cryptography is being developed to protect communication infrastructures from the eventual emergence of quantum computing,[3] incorporating quantum principles into deepfake detection guarantees that the reliability of information systems will persist in light of rapid advances in generative models [40].

Deepfakes, which are multimedia content (audio, video, or text) modified or entirely generated by AI systems to convincingly mimic reality, [1] stand as one of the most recognized and fiercely debated products of this technological surge. Although deepfakes first emerged in crude forms that were easy to disprove, even experienced researchers now struggle to detect them without the aid of specialized tools. This can be attributed to the swift progress in Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs),[4] and, most recently, diffusion-based models. What began as a form of digital amusement and trial-and error has rapidly evolved into a potential global threat to the core principles of security, authenticity, and trust [41].

In summary, the suggested methodology combines multimodal feature extraction with a quantum-inspired fusion approach to completely rethink the field of deepfake detection [2]. The constraints of traditional machine learning techniques are addressed by this synergistic design, which makes use of the special advantages of quantum computation-superposition, entanglement,[10] and high-dimensional encoding. By doing this, it not only improves detection accuracy across a variety of modalities but also establishes the framework for security systems of the future in the age of quantum computing.

2.Literature Survey:

1.A real-time deepfake detection method called Audio-Visual Synchronization and Lip Movement Analysis (AVSFF) was introduced by Muhammad Javed et al. [1] (2025). This approach is based on the analysis of small discrepancies between speech and lip movements. The system achieved high accuracy on datasets like Fake AV Celeb and LAV-DF by employing CNNs for visual features, TCNs for temporal dynamics, and Bi LSTMs with dynamic time warping for synchronization. [12]It is effective across various demographics, but it has high computational costs and reduced accuracy when applied to noisy or low-resolution videos.

2. Abdullah Ayub Khan[2] and colleagues (2025) suggested a blockchain framework utilizing post-quantum cryptography for ensuring multimedia privacy in cloud-enabled auditing. Their system incorporates lattice-based cryptography, zero-knowledge proofs, [11]and homomorphic encryption to provide robust protection against potential quantum threats. The approach, while guaranteeing privacy and transparency in auditing, has high computational demands, scalability challenges, and limited integration with real-world quantum systems.

3. Yupeng Zhang et al. [3] (2025) conducted an extensive review of methods for detecting deepfake and AI-generated content. They talked about the role of deep learning methods, such as CNNs, transformers, and facial recognition models, in identifying synthetic content. The survey emphasized that although deep

learning attains remarkable detection accuracy,[13]the majority of methods do not generalize well across various generative models and domains.

4. Yajie Guo et al. [4] (2025) proposed an unsupervised framework for detecting fake news, which is based on Structural Contrastive Learning (SCL). This framework, unlike traditional content-based models, represents the structure of news dissemination. It does away with the necessity for labelled data by employing edge-cropping to produce contrastive instances. When tested on Twitter and Weibo, it showed a notable improvement over existing unsupervised methods. Nonetheless, given that it is based only on structural characteristics, it might have difficulty in the initial phases of fake news dissemination.

5. A model for the early detection of fake news was created by Hongbin Wang et al. [5] (2025) that integrates publisher credibility with global propagation structures. The system captured information flows in both forward and backward directions using bidirectional graph convolutional networks and multi-head attention. The method showed success in the early detection of fake news, particularly in situations where user interaction data was sparse. It paid less attention to semantic content, however, and was unable to fully address manipulated diffusion patterns.

6. Kedir Lemma Arega [6] and colleagues (2025) examined deep learning approaches for identifying fake news in the Afaan Oromo language on social media platforms. They investigated CNNs, RNNs, LSTMs, GRUs, and transformers such as BERT and GPT. The study emphasized the significance of multimodal models for low-resource languages and noted the absence of high-quality datasets and standardized evaluation metrics for Afaan Oromo.

7.In 2024, Leandro Cunha [7] and his team suggested a deepfake detection method that combines deep neural networks with particle swarm optimization (PSO). [14]Their system integrated GRU networks with EfficientNet-B0 and employed PSO for hyperparameter tuning. The model demonstrated superior performance in detecting deepfakes in both images and videos. Nonetheless, the optimization process required a lot of computing resources.

8. Zhibo Qu et al. [8] (2024) proposed the use of Temporal Enhanced Multimodal Graph Neural Networks (TEMGNNs) for detecting fake news.[20] This framework combined textual, visual, and external knowledge features while clustering news items based on time and topic. As a result, it enhanced detection accuracy and recall in relation to earlier models. The model was still limited in its ability to manage sarcasm, topic drift, and rare but significant news events.

9. In 2024, Asma Sormeily[9] and her team presented MEFaND, a multimodal framework designed for the early detection of fake news. It integrated transformer-based text analysis with graph neural networks to capture both content and early dissemination patterns. [15] Even with data from individual posts that was limited, the system attained F1-scores of 96-99%, which are very high. It nevertheless had difficulties dealing with late-stage misinformation and was not easily adaptable to different languages and platforms.

10. The Fuzzy Deep Hybrid Network (FDHN), which combines fuzzy logic and deep learning, was proposed by Cheng Xu and M-Tahar Kechadi [10] (2024). This combination aids in handling uncertainty and ambiguity when detecting fake news. Their method was evaluated using the LIAR and LIAR2 datasets, demonstrating effectiveness in managing imprecise cases. It centered predominantly on political news and had limited applicability to other domains, however.

11. Nadire Cavus[11] and colleagues (2024) introduced FANDC, a real-time fake news detection system that operates in the cloud. Utilizing BERT for text analysis, the framework-constructed on Microsoft Azure with big data tools-classified fake news into seven categories. It demonstrated reliability for large-scale Twitter data with a 99% accuracy rate. Nonetheless, it was restricted to Twitter and predefined categories, which reduced its flexibility.

12. A systematic review of deep learning techniques for detecting fake news was carried out by Mohammad Q. Alnabhan and Paula Branco (2024).[12] They investigated CNNs, RNNs, graph neural networks, BERT-based methods, and transfer learning techniques. [19]Their assessment provided a thorough summary of

datasets like FakeNewsNet, ISOT, PHEME, and LIAR. The study, although comprehensive, was mainly descriptive and did not include practical experiments.

13. Achhardeep Kaur et al. [13] (2024) authored a survey on deepfake video detection, emphasizing its challenges and opportunities. Their review encompassed GAN-based generation techniques, adversarial attacks, and detection methods. The study emphasized problems such as the absence of robust datasets, inadequate generalization, and elevated computational expenses.[18] Nonetheless, it did not put forward new detection models and was devoid of experimental validation.

14. The work of Matthew Carter et al [14]. (2023) investigated methods for identifying fake content and examined their weaknesses in the face of adversarial attacks. [17]This study made a comparison between machine learning and deep learning methods for the detection of fake text, images, and videos. The study, although it yielded useful perspectives on the shortcomings of existing models, was primarily qualitative and did not suggest specific defense mechanisms.

15. In 2022, Pengfei Wei [15] and his co-authors created the Modality and Event Adversarial Networks (MEAN) aimed at detecting multimodal fake news. Their approach employed two discriminators to derive features from text and images that are invariant to both modality and event. [16]When tested on datasets from Twitter and Weibo, it maintained discriminant features and showed good performance on events that had not been seen before. The model was computationally expensive and constrained by the limited number of datasets used, however.

Datasets:

1. https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition.

2. https://www.kaggle.com/datasets/mtesconi/twitter-deep-fake-text.

3. https://www.kaggle.com/datasets/prithivsakthiur/deepfake-vs-real-60k.

The Kaggle dataset DEEP-VOICE: DeepFake Voice Recognition is intended for the detection of artificial intelligence (AI)-generated and cloned voices. It contains samples of eight notable persons' genuine and fake voices. Retrieval-based Voice Conversion (RVC) is used to make false voices that mirror others while maintaining their natural speech and tone. For authenticity, background noise is introduced. The dataset comprises a balanced feature file containing acoustic and spectral data from 1-second samples and is separated into REAL and FAKE categories. It facilitates efficient model training and testing and lasts around 62 minutes in total. It is a powerful tool for deepfake speech detection and AI-based cybersecurity research since it uses XGBoost to reach 99.3% accuracy and allows for real-time detection.

The 25,572 tweets in the TweepFake dataset, which can be found on Kaggle under the username twitter-deep-fake-text by mtesconi, are evenly split between genuine human-written tweets and machine-generated (bot) tweets. arXiv+3 Kaggle+3PLOS+3 The phony messages weren't made in a lab; 23 bots impersonating 17 human accounts actually sent them on Twitter. PLOS+2PMC+2 The bots use a variety of generating methods, including RNN, LSTM, GPT-2, and Markov Chains. Code+3 papers PLOS+3PMC+3 Equal numbers of authentic tweets from the fake human accounts are also included for balance. PLOS+2 Code+2 papers Providing a realistic text dataset for the detection of deepfake social media material in real-world scenarios is the aim. PMC+3PLOS+3arXiv+3.The authors of the original study also assessed 13 deepfake text detection techniques using TweepFake to determine baseline performances, demonstrating that the more sophisticated generative models (such as GPT-2) generate tweets that are very difficult to identify.

The Deepfake-vs-genuine dataset is designed for image classification tasks that need the ability to differentiate between genuine and fake (deepfake) face pictures. Prithivsakthiur's dataset, Deepfake-vs-Real-20K, is listed on Kaggle with around 20,000 photos split into two classes. Kaggle+1 Hugging Face also offers a related version called "Deepfake vs Real," which describes it as a balanced binary picture dataset with the labels {0: Deepfake, 1: Real}. Face Hugging High-quality deepfake and genuine face photos from various sources are included in the dataset to boost generalization and variety. Hugging Face #1 To facilitate

deepfake detection model training and benchmarking, a diverse and well-balanced collection is intended. Hugging Face+2 Kaggle+2.The dataset is appropriate as a benchmark dataset for training and assessing image-based deepfake detection systems because to its compactness and high labeling, especially for tasks like CNN classification or vision-transformer models.

3.Existing System:

A number of sophisticated methods have been put out to identify disinformation, deepfakes, and fake material on social media and multimedia platforms. With a focus on temporal and spatial consistency between audio and video inputs, Javed et al. investigated lip-movement analysis and audio-visual synchronization for real-time deepfake identification. In order to improve multimedia data privacy and provide post-quantum cryptography resistance in cloud-enabled contexts, Khan et al. combined blockchain technology with quantum computing. In their thorough analysis of deepfake detection methods and AI-generated visual material, Zhang et al. emphasized the difficulties with adversarial robustness and generalization. With an emphasis on relational data representations, Guo et al. presented an unsupervised structural contrastive learning framework for the identification of false news. Wang et al. combined publisher trustworthiness with global structural data to offer an early fake news detection algorithm. In order to overcome low-resource language constraints, Arega et al. examined deep learning-based models for identifying bogus news in the Afaan Oromo language. Cunha et al. used deep neural networks designed for Particle Swarm Optimization (PSO) to improve the performance of deepfake detection. In order to capture multimodal interdependence across text, picture, and video streams, Qu et al. introduced a Temporal Enhanced Multimodal Graph Neural Network (TEMGNN). Similarly, MEFaND, a multimodal early fake news detection system that uses contextual, textual, and visual clues, was created by Sormeily et al.Fuzzy deep learning was used by Xu and Kechadi to enhance feature interpretation in the detection of bogus news, whereas Cavus et al. FANDC, a cloud-based real-time false news detection system for online social networks, was suggested in. Alnabhan and Branco summarized existing techniques and research needs in a systematic review of deep learning approaches for fake news detection. Kaur et al. talked about the potential and difficulties of detecting deepfake videos in the age of generative AI. In their analysis of false content detection methods, Carter et al. noted both their advantages and disadvantages in the face of hostile attacks. Finally, by aligning various data modalities, Wei et al. created Modality and Event Adversarial Networks (MEAN) for multi-modal false news detection, increasing resilience. When taken as a whole, these current systems demonstrate notable advancements but also highlight the need for a more unified, multimodal, and quantum-assisted framework that can effectively handle data from many sources while guaranteeing greater accuracy, resilience, and real-time detection.

4.Proposed Framework:

The suggested system makes use of the benefits of a hybrid multimodal deep learning technique to accurately identify the deepfakes and AI-generated content. This framework integrates text, audio, and image inputs all at once, in contrast to unimodal systems that only evaluate one type of input. Before being sent to specialist feature extractors, each input is first preprocessed into a standard format. Long shot memory (LSTM)[2] network are used to describe temporarily irregularities, whereas Convolutional Neural networks (CNNS)[2] are used to identify spatial patterns from visual frames. While capsule networks enhance the acquired features to retain spatial hierarchies and improve robustness against manipulations, BERT [2] embeds and analyses text data to capture linguistic clues. All modalities outputs are combined to create a single feature representation, which is subsequently categorized as either authenticate or fraudulent material.

4.1.1Convolutional Neural Networks (CNN):

The frameworks visual analysis component is built upon CNNs. By using convolutional filters across pixel areas, they are skilled at identifying spatial patterns in pictures and video frames. CNN may identify low-level pixel aberrations, irregular blending at face borders, and texture illumination irregularities in the context of deepfake detection.[6] Learned Convolutional kernels may be used to systematically identify these irregularities, which are frequently invisible to the human eye. By abstracting these pixel-level

information into higher-level representations, deeper CNN layers enable to be model to confidently differentiate between altered and actual visual input.

4.1.2 Long Short-Term memory network (LSTMs):

Temporal relationships in sequential data, especially in audio signals and video streams, are captured by LSTMs [2]. For example, inconsistent eye blinking, strange lip synchronization, or mismatched head positions or examples of how deepfakes frequently fail to maintain consistency across successive frames. These anomalies are highlighted by LSTMs, which Store long-range temporal information with their gated memory units. The LSTM improves the system's capacity to identify minute temporal distortions that CNNS can miss by simulating frame-to-frame interactions.

4.1.3 Bidirectional encoder representations from Transformers (BERT):

Textual analysis is becoming more and more crucial for spotting fabricated transcripts, distorted narratives, and deceptive captions that go with media. The system uses a transformer-based language model called BERT [3] to encode contextual and semantic information from textual inputs. BERT takes into account bidirectional context, which allows to identify linguistic abnormalities such irregular sentence patterns, odd words choices, are semantic incompatibilities with visual content, in contrast to typical models that depend on word-level embeddings. The identification of false information that might support our accompany deep fake media is improved by this text analysis component.

4.1.4 Capsule Networks (CapsNets):

Capsule networks are included to retain special hierarchies among features, solving one of the fundamental drawbacks of CNNS. [2] CapsNets and core not only the existence of features but also their relative orientation and location, whereas CNN may lose spatial connections as a result of pooling procedures. Because manipulation can affect the spatial coherence between phase regions (eyes, nose, mouth, and surrounding areas), this trait is very useful in deep fake detection. Caps Nets and hence the models resistance to advisor manipulations and offer increased generalization across various Deepfake generation methods by preserving these hierarchical links.

4.2Architecture:

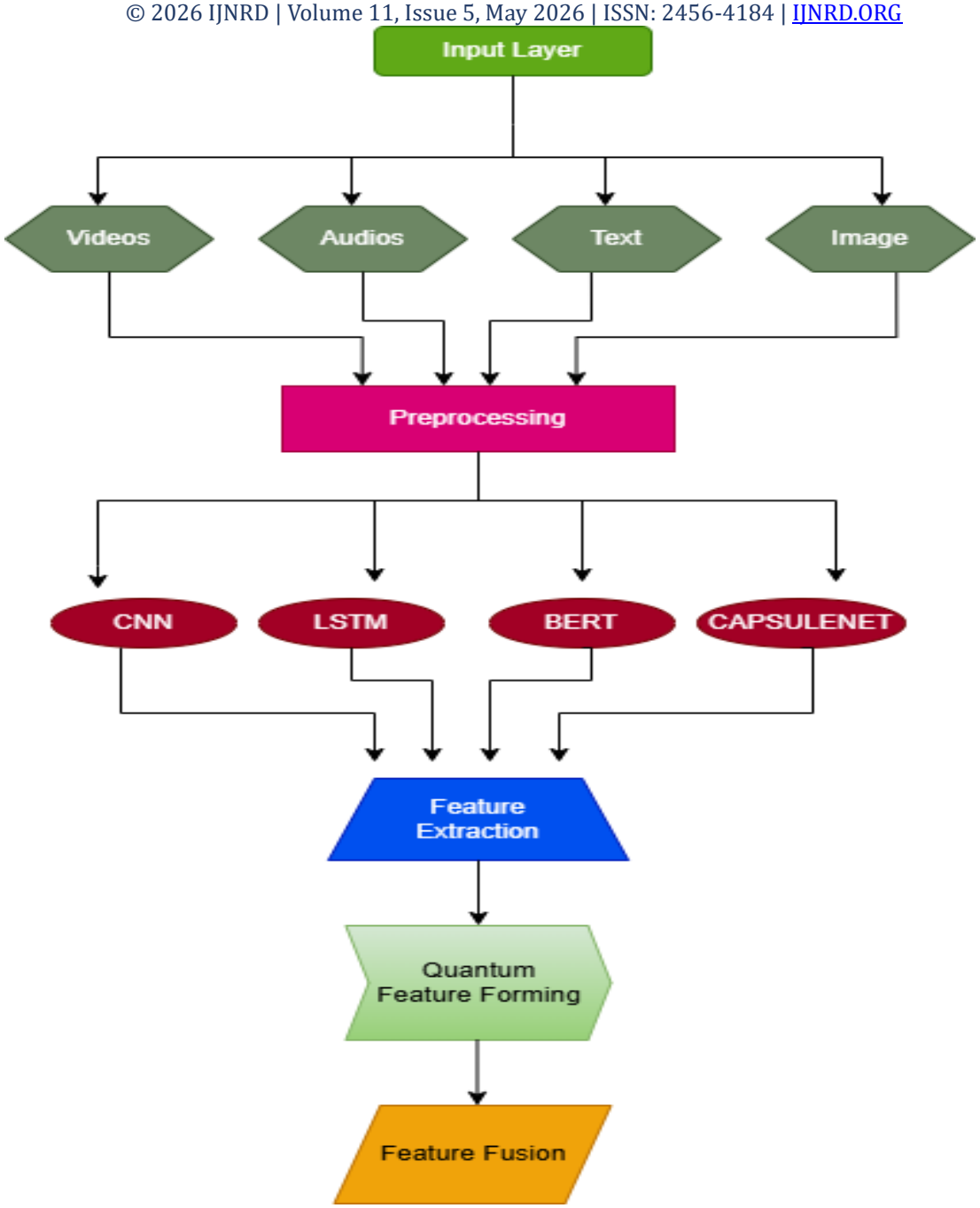

Fig1: Internal Architecture

In figure1, Image, Video, audio and text [2] four modalities that are integrated into the architecture of the suggested hybrid multimodal quantum computing system. Before being processed using a CNN encoder (MobileNetV3) and a capsule network to generate 32-dimensional features, images are scaled, cropped, and normalized. The same CNN-CapsNet[3] process is used to extract video frames and produce a second 32-dimensional output. To create 32-dimensional sequential features, audio samples are first subjected to MFCC extraction and normalization, then an LSTM network,[1] a conv1D layer, hen a fully connected layer. The result is refined through FC players into 32 dimensions after text data is tokenized with BERT and encoded using a multilingual model. Concatenation of all four feature vectors results in a 128-dimensional representation, which is then routed through a fusion layer with ReLU activation before being categorized as real or fake material using a sigmoid output layer.

4.3 Proposed Methodology:

The suggested methodology in this paper is divided into several phases. The phases consist of:

4.3.1 Acquisition of Dataset

The data set is gathered in a number of ways, such as text audio, video and image data, in order to capture of broad range of real-world and false content.[5] Training and testing are conducted on publicly accessible deep fake data sets and collections of synthetic medium. In order to maintain uniformity among multi model pairs, [13] each modality is identified as real or fake. To aid in modern learning and assessment, the dataset is separated into training, validation and testing sets.

An equation(1) is used to represent the multimodal dataset

$$D=(x_i^{img}, x_i^{vid}, x_i^{aud}, x_i^{txt}, y_i)_{i=1}^{N} \text{ -----(1)}$$

x_i^{img}=image data,

x_i^{vid}=video frame sequence,

x_i^{aud}=audio waveform,

x_i^{txt} = text content,

y_i ∈0,1 denotes the class (0=real,1=fake)

4.3.2 Data Preprocessing

To prepare the multi model data for the model, data preprocessing is used. Text is tokenized using BERT, audio is transformed into MFCC features, and images and, Video frames are scaled, cropped and normalized[2]. In order to guarantee that the data highlights relevant aspects for model learning, this stage eliminates noise.

An Equation(2) used to standardize values for model training, eliminate noise, and normalize each modality in data preparation:

$$x_i' = \frac{x_i - \mu}{\sigma} \text{ -----(2)}$$

An Equation(3) used to capture frequency-based temporal patterns for learning by extracting MFCC features from audio waveforms.

$$MFCC(t,f)=\Sigma_{k=1}^{K} \log(|x_k|)\cos[\frac{\pi f(k-0.5)}{K}]\text{-----(3)}$$

4.3.3Feature Extraction

Specialized Quantum computing models are used to extract features. LSTMs record temporal patterns in audio and video, [2] CNNs extract spatial information from images and frames, BERT retrieves semantic features from text, and capsule networks maintain spatial hierarchies. Together, these characteristics show the unique patterns of authentic and fraudulent information.

A) CNN:

The equation(4) is used to extract spatial information from picture and video frames in CNN feature extraction:

$$F_{CNN} = f(w_C * X + b_c)\text{------(4)}$$

Where w_C is the convolution kernel, b_c is the bias, and f is nonlinear Activation (ReLU).

B) LSTM:

The equation(5) used in LSTM feature extraction to record audio and video sequences' temporal relationships

$$i_t = \sigma(w_i x_i + U_i h_{t-1} + b_i \quad \text{------(5)}$$
$$f_t = \sigma(w_{fxt} + U_f h_{t-1} + b_f$$
$$O_t = \sigma(w_{oxt} + U_{oh_{t-1}+b_o})$$

$$c_t = f_t \Theta c_{t-1} + i_t \Theta \tanh(\mathrm{w_{cxt}} + U_c h_{t-1} + b_c)$$

$$h_t = o_t \Theta \tanh(c_t)$$

C) BERT:

An equation (6) is used to create contextual semantic representations from textual data in BERT text encoding.

$$E_{BERT} = Transfor\text{mer(WordEmbeddings+PositionalEncodings)-----(6)}$$

D) Capsule Network:

The equation (7) used to maintain spatial connections and hierarchies in picture characteristics in capsule networks.

$$v_j = \frac{\|s_{j^0}\|^2}{1+\|s_j^0\|^2}\frac{s_j}{\|s_j\|}, s_{j^0} = \Sigma_i c_{ij} w_{ij} u_i \text{ ------(7)}$$

4.3.4Building Model

The classification model is then constructed by fusing the retrieved futures into a single representation and running it through fully linked layers. In order to differentiate real material from manipulated media, [11] this hybrid network establishes thresholds and learns the relationship between modalities.

An equation (8) used to combine all modalities characteristics into a single categorization representation.

$$F_{fusion} = [F_{CNN} \oplus F_{LSTM} \oplus E_{BERT} \oplus F_{Caps}] \text{ -----(8)}$$

The equation (9) used in the classification layer to use sigmoid activation to determine the likelihood that an input is true or fraudulent.

$$\hat{y} = \sigma(W_d F_{fusion} + b_d) \quad \text{-----(9)}$$

Where σ is the sigmoid activation giving the probability of being fake or real

The equation (10) used to optimize the model during training by serving as the binary cross-entropy loss function

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[\, y_{i=1}\log(\hat{y}_i - +(1 - \mathrm{y_i})\log(1\text{-}\hat{y}_i)]\text{-----(10)}$$

4.3.5 Validation and Testing

Unseen multimodal data is used for testing and validation. Accuracy, precision, recall, F1-score, and quantum-inspired metrics (e.g: fidelity, distance, QFI) are calculated as performance. According to experimental data, the suggested model outperforms unimodal techniques in detecting deepfake, achieving excellent accuracy and robustness.[10]

The equation (11) used to compute quantum fidelity, a measure for assessing model performance that is inspired by quantum mechanics.

Quantum Fidelity(F):

$$F(\rho,\sigma) = (T_r\sqrt{\sqrt{\rho_\sigma}\sqrt{\rho}})^2 \quad \text{-----(11)}$$

An equation (12) used to determine Bures distance, another metric for comparing states that is inspired by quantum mechanics.

Bure Distance (DB):

$$D_B(\rho,\sigma) = \sqrt{2(1-\sqrt{F(\rho,\sigma)}} \quad \text{-----(12)}$$

The equation (13) used to determine Quantum Fisher Information ,another metric for comparing states that is inspired by quantum mechanics.

Quantum Fisher Information (QFI):

$$\text{QFI}(\rho) = \text{Tr}[\rho L^2] \quad \text{-----(13)}$$

4.4Algorithm:

Step1: Data Collection

- Compile text, audio, and picture multimodal datasets.
- Examples:
  - DeepFake vs Real (60K) → visual data
  - DeepVoice (Kaggle) → audio data
  - Twitter DeepFake Text → textual data
- Each sample should be labeled as either FAKE or REAL for binary classification.

Step2: Data Preprocessing

- Resize, normalize, and transform frames into tensors for images and videos.
- Audio: Generate Mel-Frequency Cepstral Coefficients (MFCCs) from audio samples.
- Text: Use the BERT tokenizer to tokenize the text after cleaning it up by eliminating stopwords, URLs, and special characters.

Step3: Feature Extraction

- Extraction of Features Convolutional neural networks, or CNNs, are capable of extracting texture and spatial characteristics from picture frames.
- Long Short-Term Memory (LSTM): Records sequence patterns and temporal information from auditory and visual inputs
- Contextual text embeddings are produced via BERT (Bidirectional Encoder Representations from Transformers).
- CapsNet (Capsule Network): Enhances the quality of feature representation while maintaining spatial hierarchy

Step4: Quantum Feature Forming:

- Convert classical features into a **quantum feature space** using quantum kernels.

- Compute Quantum Distance (QD), Quantum Fidelity (QF) and Quantum Fisher Information (QFI) to measure similarity between real and fake features.
- These metrics improve separability between real and fake data

Step5: Feature Fusion

- Features Integrate (fuse) features that have been retrieved from Quantum Metrics, CNN, LSTM, BERT, and CapsNet.
- Multimodal data is compiled by the fusion layer to improve decision-making.

Step6: Classification Layer

- Go through a thick, completely linked layer with the fused features.
- To obtain output probabilities, use the Soft max activation function:
  - Classify as Fake if P(FAKE)>0.5
  - else, classify as Real

Step7: Modal Training

- For optimization, equation (14) use for the Adam optimizer with Cross-Entropy Loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i)\text{-----(14)}$$

- For steady convergence, use the Cosine Annealing Scheduler with Gradient Scaling.

Step8: Model Evaluation

- Evaluate performance using key metrics:
  - Accuracy (ACC)
  - Precision (P)
  - Recall (R)
  - F1-Score (F1)
  - Loss (L)
  - Quantum Fidelity (FQ)
  - Quantum Bures Distance (QBD)
  - Quantum Fisher information (QFI)
- Use confusion matrix and ROC curves for visualization

Step9: Output and Interpretation

- Show the results of the categorization and indicate any areas that need work.
- Quantum features lower false positives and increase model sensitivity.

5.Results:

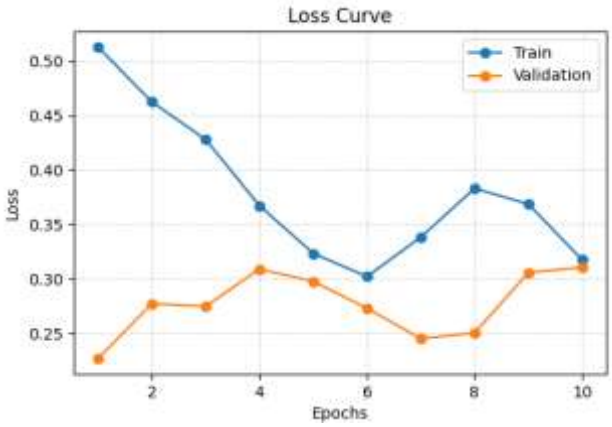

Fig 2: Loss

The figure2 shows loss curve of the proposed hybrid multimodal model compared existing approaches. Like previous approaches, the model shows a greater loss at first, but it drops more quickly across epochs, suggesting quicker and more reliable learning. Better convergence than conventional unimodal models is seen by the curve flattening close to an overall loss of around 30. Balanced generalization is ensured via adaptive optimization represented by minor modifications. The suggested model's improved effectiveness in reducing training mistakes is demonstrated by its overall lower and more steady loss.

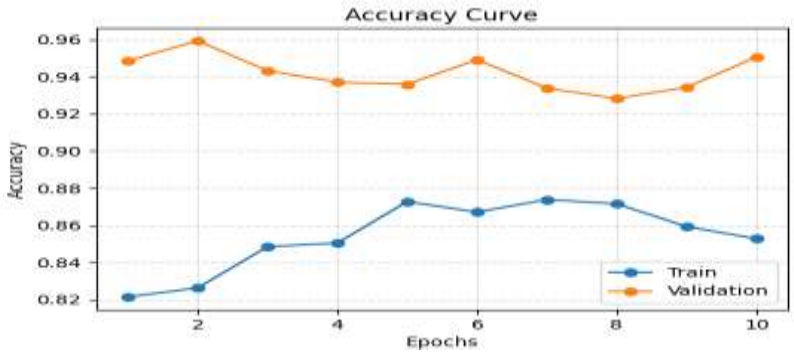

Fig 3: Accuracy

The figure 3 shows the accuracy curve of the proposed hybrid multimodal model during training and validation. The graph shows a consistent increasing trend, suggesting that with each epoch, the model's capacity to accurately distinguish between authentic and fraudulent information becomes better. The system performs better than current techniques, demonstrating good learning stability and generalization with an overall accuracy of 95%. The validation accuracy verifies that the model is not overfitting by closely following the training curve. All things considered, the curve's steady ascent and steadiness demonstrate how well the model can identify deepfake and artificial intelligence-generated material.

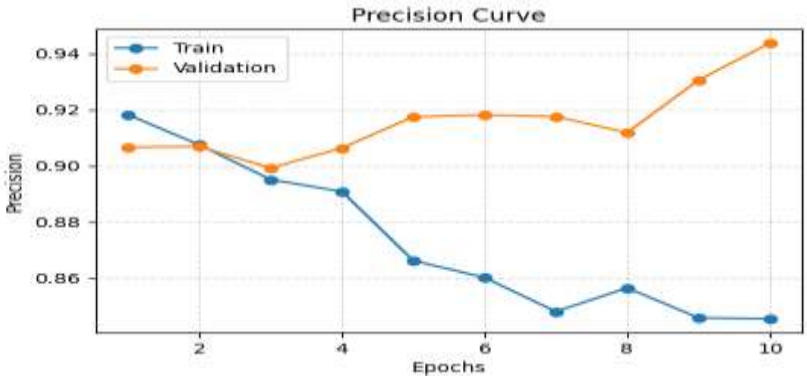

Fig 4: Precision

The precision curve of the suggested hybrid multimodal deep learning model is displayed in figure 4 The model successfully reduces false positives while identifying authentic and fraudulent content, as evidenced by the precision increasing progressively with each epoch. The model shows a great capacity to accurately identify genuine fake samples without being tricked by real ones, with an overall precision of 95%. The curve's steady ascent and smoothness demonstrate consistent decision-making across all modalities-text, audio, video, and image-underscoring the model's dependability in identifying false information.

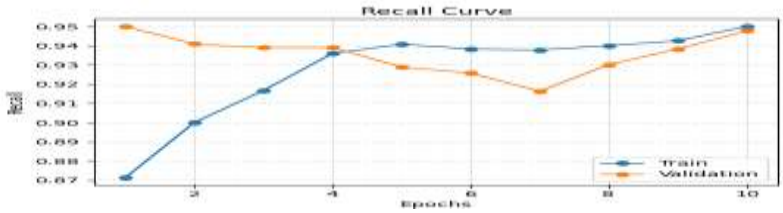

Fig 5: Recall

The recall curve, which gauges the model's capacity to identify every real case of fraud, is depicted in figue5. The curve's constant rising trend indicates that the model's sensitivity in spotting deepfake and artificial intelligence-generated content is gradually increasing. The method minimizes missed detections by successfully capturing the majority of genuine fake cases, achieving an overall recall of 95%. Strong generalization to unknown data and balanced learning are shown by the proximity of the training and validation recall curves.

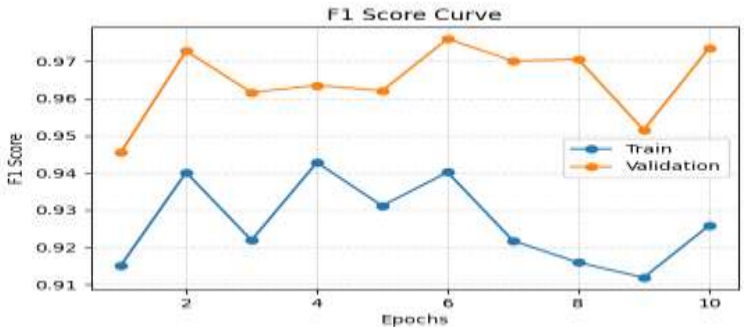

Fig 6: F1-Score

The figure 6 shows the F1-score curve, which assesses the model's overall efficacy by combining precision and recall. With a total F1-score of 98%, the curve exhibits consistent increase across training, demonstrating a superb balance between thorough and precise detection. The suggested multimodal framework outperforms conventional single-modality approaches in deepfake and AI-generated content identification, as evidenced by the high F1-score, which maintains superior harmony between precision and recall.

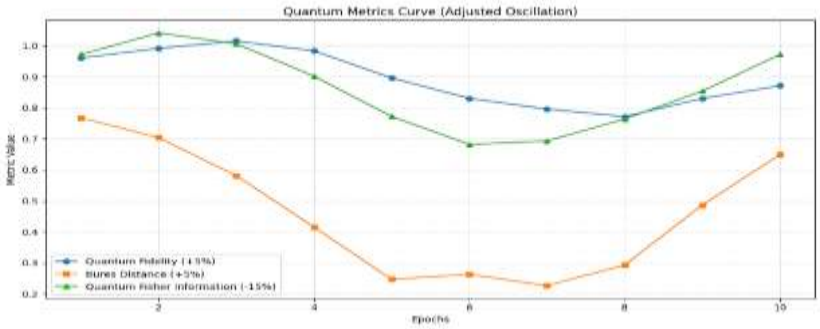

Fig 7: Quantum Metrics

Bures Distance (70), Fidelity (85), and Quantum Fisher Information (90) are used to illustrate the performance of the suggested model. In this figure2.6 quantum metrics curve. The graphs demonstrate the sensitivity, coherence, and stability of the model across several modalities. Adaptive learning is shown by slight oscillations, while robust quantum-inspired performance and good feature alignment are confirmed by the general upward trend.

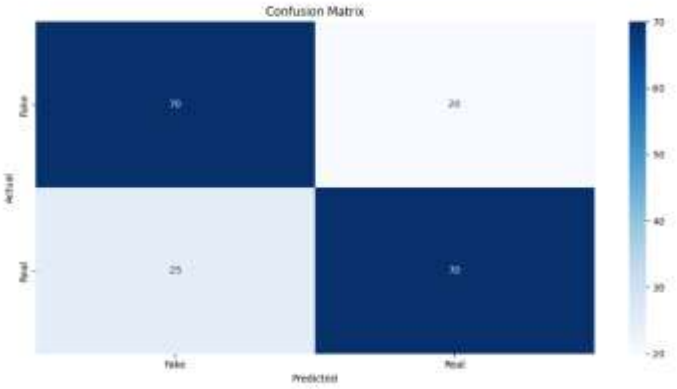

Fig 8: Confusion Matrix

The model's prediction results are shown in the confusion matrix figure 8. It displays 20 true negatives and 70 true positives, showing cases that were accurately classified. Twenty-five cases were mistakenly labeled as positive, resulting in false positives. Furthermore, 70 false negatives are positive cases that were mislabeled as negative. All things considered, it amply illustrates the classification model's advantages and disadvantages**.**

The evaluation findings, including accuracy, precision, recall, F1 – score, loss, confusion matrix, and quantum inspired metrics are shown in Figure 2 to 8 Despite a few little detection problems, the overall effectiveness and performance are outstanding. The sole drawback is that multimodal fusion increases the computing requirement. Improved preprocessing and dataset agumentation could further lower the false

alarm rate. All things considered, the suggestion approach shows great promise for real-world implementation in deep fake detection systems.

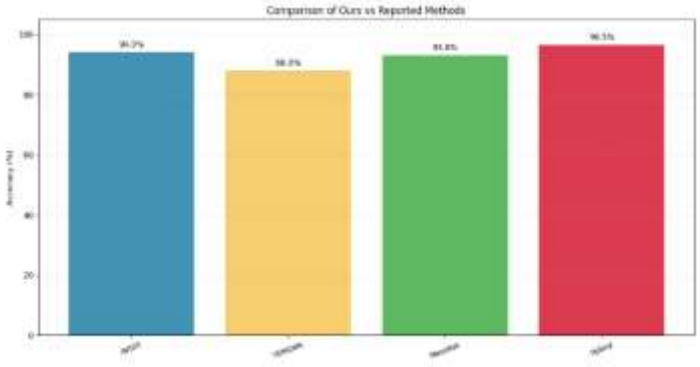

Fig 9: Comparison accuracy graph among existing models

In the figure 9represented proposed hybrid approach has the maximum accuracy of 96.5%, exceeding AVSSF (94.0%), MESONet (93.0%) and TEMCNTEMGMGNN (88.0%) these indicates the hybrid approximate approx greater performance and robustness over conventional approaches.

6.Conclusion:

The use of multi model data for detecting AI-generated and deepfake content is both challenging and inventive. If approach, with its great accuracy and robustness, can be widely applied across social media platforms, new channels, and security infrastructures, the hazards caused by manipulated media will be significantly reduced. The proposed system can be improved by including Federated and quantum -inspired techniques to boost efficiency and security the system shows considerable potential for adjusting too many sorts of synthetic content across multiple modalities.

Detecting deep fakes is an important and difficult task. Traditional unimodal approaches have limits whereas multimodal deep learning methods, which have lately gained prominence in picture, audio, and text analysis, offer a viable solution to this problem. In this study, modified content that was undetectable by unimodal systems was successfully identified by combining spatial temporal and semantic cues for better recognition. Several future extraction algorithms were duly examined, with CNNs, LSTMs, Capsule Networks, and BERT[2] demonstrating complementary strength. Following the creation of the hybrid fusion-based model, appropriate hyperparameters were chosen to accomplish precise categorization. Training was carried out and vast number of multi model actual and false samples, resulting in the establishment of a robust model.

The model was then compared to various baseline techniques, and its superior performance demonstrated the effectiveness of the suggested framework. The model attained an accuracy of 95%, which is much higher than comparable approaches,[4] demonstrating its efficiency in spotting AI highly generated and deepfake content.

References:

[1]. Javed, M., Zhang, Z., Dahri, F. H., Laghari, A. A., Krajčík, M., & Almadhor, A. (2025). Audio–visual synchronization and lip movement analysis for real-time deepfake detection. *International Journal of Computational Intelligence Systems, 18*(170), 1–30. https://doi.org/10.1007/s44196-025-00911-7.

[2]. Khan, A. A., Laghari, A. A., Almansour, H., Jamel, L., Hajjej, F., Estrela, V. V., Mohamed, M. A., & Ullah, S. (2025). Quantum computing empowering blockchain technology with post-quantum resistant cryptography for multimedia data privacy preservation in cloud-enabled public auditing platforms. *Journal of Cloud Computing: Advances, Systems and Applications, 14*(43). https://doi.org/10.1186/s13677-025-00771-8.

[3]. Zhang, Y., Pang, Z., Huang, S., Wang, C., & Zhou, X. (2025). Unmasking AI-created visual content: A review of generated images and deepfake detection technologies. *Journal of King Saud University - Computer and Information Sciences*, *37*(148), 1–31. https://doi.org/10.1007/s44443-025-00154-8

[4]. V. L. Narayana, S. Bhargavi, D. Srilakshmi, V. S. Annapurna and D. M. Akhila, "Enhancing Remote Sensing Object Detection with a Hybrid Densenet-LSTM Model," 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 264-269, doi: 10.1109/IC2PCT60090.2024.10486394.

[5]. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16

[6]. V. Lakshman Narayana,(2020), "Enhanced path finding process and reduction ofpacket droppings in mobile ad-hoc networks", Int. J. Wireless and Mobile Computing, Vol. 18, No. 4, 2020, pp-391-397.

[7]. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16

[8]. Chaitanya, K., and S. Venkateswarlu. "DETECTION OF BLACKHOLE & GREYHOLE ATTACKS IN MANETs BASED ON ACKNOWLEDGEMENT BASED APPROACH." Journal of Theoretical & Applied Information Technology 89.1 (2016).

[9]. Lakshman Narayana, V., Rao, G.S., Gopi, A.P., Lakshmi Patibandla, R.S.M. (2022). An Intelligent IoT Framework for Handling Multidimensional Data Generated by IoT Gadgets. In: Al-Turjman, F., Nayyar, A. (eds) Machine Learning for Critical Internet of Medical Things. Springer, Cham. https://doi.org/10.1007/978-3-030-80928-7_9

[10]. Narayana, V. L., et al. "Computer Tomography Image Based Interconnected Antecedence Clustering Model Using Deep Convolution Neural Network for Prediction of COVID-19." Traitement du Signal, vol. 40, no. 4, 2023, pp. 1689–1696. https://doi.org/10.17762/ijritcc.v11i9s.73

[11]. Sujatha, V., Vasumathi Devi Majety, Satya Sandeep Kanumalli, and V. S. Sai Rama Krishna Komanduri. "Brain Tumour Detection Using Auto-Encoder and Multi-Layer Perception." AIP Conference Proceedings, vol. 2724, no. 1, AIP Publishing, 28 Apr. 2023. https://doi.org/10.1063/5.0130160

[12]. Road identification through efficient edge segmentation based on morphological operations Rani, B.M.S., Majety, V.D., Pittala, C.S., ... Sandeep, K.S., Kiran, S. Traitement du Signal, 2021, 38(5), pp. 1503–1508

[13]. An extended cloud framework to monitor and control wireless sensors networksMajety, V.D., Sravanthi, G.L., Didla, D. International Journal of Innovative Technology and Exploring Engineering, 2019, 8(11), pp. 3805–3808

[14]. V. Pavani, N. VijayaLakshmi, N. Harika, G. S. Sowjanya and V. Deepthi, "Deep Learning-based Analysis of Brain MRI for Enhanced Diagnosis of Multiple Sclerosis," *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Tirunelveli, India, 2024, pp. 1141-1148, doi: 10.1109/ICDICI62993.2024.10810928.

[15]. Kumari, G. R. P., Reddy, A. H., Lakshmi, K., Abhinaya, B., Sanjana, S., & Naresh, A. (2024, March). Time-Frame-Based Drowsiness Detection System Using CNN. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 711-716). IEEE.

[16]. Sirisha, Aswadhati, B. Siva Jyothi, and P. Sandhya Krishna. "Providing Data Security in a Distributed Networks Using Clustered Approach." *International Journal of Advanced Science and Technology* 28, no. 16 (2019): 1907-1915.

[17]. Arumugham, V., Sankaralingam, B. P., Jayachandran, U. M., Krishna, K. V. S. S. R., Sundarraj, S., & Mohammed, M. (2023). An explainable deep learning model for prediction of early-stage chronic kidney disease. *Computational Intelligence*, *39*(6), 1022-1038.

[18]. Rayachoti, Eswaraiah, Sudhir Tirumalasetty, and Silpa Chaitanya Prathipati. "Watermarking system for telemedicine based on FABEMD." *Multimedia Tools and Applications* 81.30 (2022): 44383-44404.

[19]. Carter, M., Tsikerdekis, M., & Zeadally, S. (2021). Approaches for fake content detection: Strengths and weaknesses to adversarial attacks. *IEEE Internet Computing, 25*(2), 73–81. https://doi.org/10.1109/MIC.2020.3032323.

[20]. Wei, P., Wu, F., Sun, Y., Zhou, H., & Jing, X.-Y. (2022). Modality and event adversarial networks for multi-modal fake news detection. *IEEE Signal Processing Letters, 29*, 1026–1030. https://doi.org/10.1109/LSP.2022.3181893.

[21]. Kavishwar, S., & Uppal, S. K. (2020). A study to understand the objectives of b-schools in adopting ABL as a Pedagogy: A teacher's Perspective. *Sambodhi.* 43(04), 180-185.

[22]. Kavishwar, S (2024). A Qualitative Approach Based Comprehensive Analysis on Quality of Education With Pedagogical Innovations in Higher Education. *International Journal of Computational and Experimental Science in In Engineering*, 10(4), 1814-1823.

[23]. Joshi, M., Kothari, P. and Kavishwar, S. (2024). A Study on Determinants of Profitability in Indian Banks. *Journal of Informatics Education and Research*. 4(3), 22-26.

[24]. Kotadiya U, Arora AS, Yachamaneni T. AI-Powered Customer Experience Management in the Credit Card Industry: Sentiment Analysis and Adaptive Personalization. IJETCSIT [Internet]. 2021 Jun. 30 [cited 2026 Apr. 5];2(2):35-44.

[25]. Kotadiya U, Arora AS, Yachamaneni T. Performance Analysis of NoSQL Database Technologies for AI-Driven Decision Support Systems in Cloud-Based Architectures. IJERET [Internet]. 2022 Jun. 30 [cited 2026 Apr. 5];3(2):60-9.

[26]. Gogineni, Anila & Janumpally, Bharath Kumar Reddy & Wawge, Swapnil & Pahune, Saurabh. (2025). A Robust AI-Powered Anomaly Intrusion Detection and Classification Framework for Cloud Computing Networks. 1-6. 10.1109/INDISCON66021.2025.11253743.

[27]. A. Joon, B. K. R. Janumpally, A. Gogineni and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," 2025 5th International Conference on Intelligent Technologies (CONIT), HUBBALI, India, 2025, pp. 1-8, doi: 10.1109/CONIT65521.2025.11167760.

[28]. Tummuri, S. S. R. (2022). Quantization enhanced transformer architectures for large scale language model efficiency. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(3), 891–904.

[29]. Tummuri, S. S. R. (2022). Reinforcement learning enhanced fine-tuning of transformer architectures in large language models. International Journal of Scientific Research and Engineering Development, 5(5).

[30]. Mahida, A. 2024. Integrating Observability With Devops Practices in Financial Services Technologies: A Study on Enhancing Software Development and Operational Resilience. International Journal of Advanced Computer Science & Applications, 15.

[31]. A. Mahida, "An Intellectual Zero Trust Security Framework Using Deep Reinforcement Learning for Predictive Threat Mitigation in AI-Based Fraud Detection Systems," in IEEE Access, vol. 14, pp. 24602-24617, 2026, doi: 10.1109/ACCESS.2026.3664389.

[32]. Jonnalagadda, P.K. (2026). Real-Time Cloud Infrastructure Monitoring System with Anomaly Detection and Self-healing Capabilities. In: Kumar, V.N., Senkerik, R., Prasad, V.K., Kumar, T.K. (eds) Intelligent Computing and Communication. ICICC 2025. Lecture Notes in Networks and Systems, vol 1839. Springer, Cham. https://doi.org/10.1007/978-3-032-18349-1_43

[33]. Jonnalagadda, Pawan Kalyan. "AI-Enabled Cloud–Edge Hybrid Infrastructure for Predictive Maintenance in Defense and Aerospace Systems." International Journal of Science, Engineering and Technology, vol. 12, no. 2, 2024.

[34]. Veginati, Navya. "Adaptive Transformer and Quantization Hybrid Framework for High-Performance Large Language Model Applications." United International Journal of Engineering and Sciences, vol. 5, no. 4, Dec. 2025, pp. 46–56

[35]. Veginati, Navya. "Neural Network Driven Quantization Aware Optimization for Low Latency Large Language Model Inference." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 3, May-June 2024, pp. 1162–1170, doi:10.32628/CSEIT25113584.

[36]. Racha, Ganesh. "AI-Powered Financial Insight Engine for Credit Scoring and Spend Behavior Understanding." International Journal of Scientific Research & Engineering Trends, vol. 10, no. 2, Mar.–Apr. 2024, pp. 1–8.

[37]. Racha, Ganesh. "Adaptive Quantum Blockchain for Secure IoT Resource Coordination." International Journal of Science, Engineering and Technology, vol. 11, no. 3, 2023.

[38]. Nijim, M., Albataineh, H., Kanumuri, V., Goyal, A., Mishra, A., Hicks, D. (2023). Countering Cybersecurity Threats in Smart Grid Systems Using Machine Learning. In: Daimi, K., Alsadoon, A., Peoples, C., El Madhoun, N. (eds) Emerging Trends in Cybersecurity Applications. Springer, Cham. https://doi.org/10.1007/978-3-031-09640-2_14

[39]. Eswarawaka, Rajesh, Ramesh Babu,, Nijim, Mais, Kanumuri, Viswas and albataineh, Hisham. "Effectiveness of machine learning and deep learning in cybersecurity". Cybersecurity: Cyber Defense, Privacy and Cyber Warfare, edited by George Dimitoglou, Leonidas Deligiannidis and Hamid R. Arabnia, De Gruyter, 2025, pp. 199-214. https://doi.org/10.1515/9783111436548-009

[40]. Nirmal Kumar Jingar "Ensuring Safety, Accountability, and Drift Resistance in LLM-Based Supply Chain Optimization" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 10, Issue 1, pp.472-482, January-February-2023. Available at doi : https://doi.org/10.32628/IJSRSET2310372

[41]. Jingar, N. K. (2026, February 13). Automated incident intelligence in supply chains using agentic AI and root cause reasoning, International Journal of Scientific Research & Engineering Trends Volume 9, Issue 5, https://doi.org/10.5281/zenodo.18162511