Volume 10, Issue 11, November – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov797

Text Summarization Using LLM

Utsha Sarker[1]; Lalit Vaishnav[2]; Archy Biswas[3]; Ashish Raj[4]; Saurabh[5]

[1,2,3,4,5] Department of AIT-CSE
Apex Institute of Technology
Chandigarh University, Punjab, India

Publication Date: 2025/11/24

Abstract: The main reason for the high effectiveness of text summarization is due to the success of LLMs for this task and across different domains. This work aims at understanding how LLMs are used to summarize domains and make it more accurate and efficient. We discuss how current models perform with regard to specialized information, with focus on the financial and medical domains. The work suggests that an approach using Vertex AI, a generative machine learning platform in the cloud, can be used to assess pre-trained summarization models for different tasks. Most of the research presented in the paper also reveals the efficacy of Vertex AI for text summarization with high accuracy and efficiency. We demonstrate the applicability of the platform for summarizing transcripts and dialogues, generating bullet points, titles and to- do lists. Also, the research show that Vertex AI is reliable in terms of cost since it can be used by businesses and individual researchers.

Keywords: *LLM, Summarization, Domain-Specific, Vertex AI, Generative Models, ML, NLP, Finance, Healthcare, Evaluation, ROUGE, Cloud-Based, AI, Data Science, Text Mining.*

How to Cite: Utsha Sarker; Lalit Vaishnav; Archy Biswas; Ashish Raj; Saurabh (2025) Text Summarization Using LLM. *International Journal of Innovative Science and Research Technology*, 10(11), 1193-1198. https://doi.org/10.38124/ijisrt/25nov797

I. INTRODUCTION

The unstoppable expansion of computerized processes and techniques has produced an overwhelming flow of textual data where to look for relevant information. Text summarization, a process of converting long documents into shorter summaries as a means of dealing with information overload, is a technique used most frequently at the present time. Previous work on summarization involves the use of statistics to count words, or rule based approaches which do not approach deep problems in the text when the text is sophisticated and at a knowledge level in areas such as finance or healthcare [1].

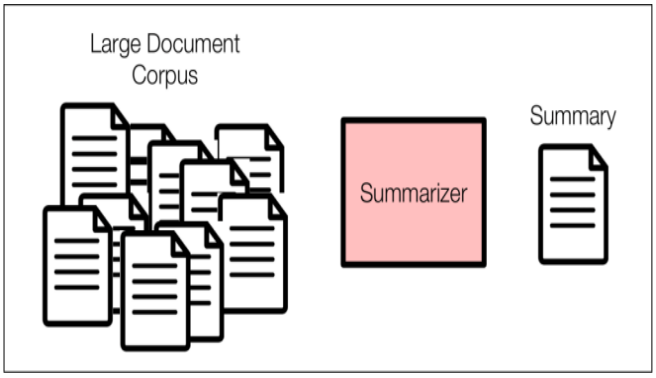

Fig 1. Representing the Need for the Text Summarization Tool

Tailored AI has been used in text summarization for some time, but recent developments such as the Large Language Models (LLMs) have made this even better. These feedback models are robust models based on huge data sets and can produce accurate and relevant summaries more often than not which are also faster than manually written summaries. Nevertheless, several issues can be still observed as regards specific adoption of the LLMs for particular domains, including ability to properly handle specific features of certain text types, like financial reports, or to properly capture and analyze the details of certain medical terminology [3].

To overcome these challenges, this research explores the possibility of Vertex AI, a cloud-based powerful machine learning service provided by Google Cloud. Vertex AI is a platform that offers tools and services for building, training and deploying of machine learning models [4]. In this study, we plan to extend the functionalities offered by generative models of Vertex AI to analyze how well and on what basis text summarization tasks can be performed within the fields of finance and healthcare. The study aims to:

- Evaluate the performance of LLMs in summarizing domain-specific text data.
- Explore the feasibility of Vertex AI for automating text summarization tasks.
- Analyze the cost-effectiveness and scalability of Vertex AI for real-world applications.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov797

This work truly marries the real needs of domain-based apps on Vertex AI with the wide summarizing skill of the big language model. Here, a more focused method has been taken to handle hard needs such as financial and medical texts-meanings of tough, special terms, proper sense of context, and keeping care for exactness inside key information limits [5].

Such difficult tasks of serious areas say a lot about the usefulness and credibility of the text summary tool. This work tries to study in greater detail the flexibility of many generative models presently being improved on Vertex AI for matching the special features of such expert fields. Financial writings are normally filled with statistical details, references to legal rules, and business plans, while complex medical words, clinical instructions, and facts for patients are included in healthcare writings [6]. The learning from how well Vertex AI fares in such cases will provide a clear view about its power to raise up work speed and assist in making better choices inside expert environments. Besides performance number checking, this study describes the growth and size-handling issues of such advanced systems. While choosing AI-based summary tools, the major worries for a company are how well they can grow and at what cost of owning them. It would be verified whether the usage of Vertex AI could handle big data needs in a quick and inexpensive way for every kind of organization [7].

In this way, the study desires to give strong support to those areas of summarization that are made by AI and change every day. The results will guide the improvement not just in the strength of Vertex AI but will also help in building better AI summary tools in the future, mainly in those areas where meaning correctness and expert knowledge are needed most. These results may start new paths for using AI to fully change how companies read, use, and get value from written data.

II. LITERATURE SURVEY

Quite a number of works have been done on LLMs for summarization, majorly in special fields like finance and healthcare. Various approaches have been employed toward improving the correctness and value of summaries, mostly for handling complex, field-based information. Such growth has been driven by how a human understands the situation, reads language in a given field, and finds meaning in a lot of unorganized writing. These models, at their deep learning architecture, were constructed to handle large data and then produce clear, smooth summaries. Previous reviews have demonstrated that models are constantly improved for better contextual understanding and solving problems of summarization for domain-based information, which remains one of the most important future study directions.

Ezhilian Wilson and team [8] introduced the new setup to improve financial text summarization, FIN2SUM. The current work explained the key parts of 10-K reports from top NASDAQ-listed firms, mainly the management's talk and analysis areas. FIN2SUM was used as a setup to test Large Language Models, mostly the skills of LLAMA-2 for hard financial writing. The paper by Wilson gave a comparison of models like LLAMA-2, FLAN, and Claude 2, tested with BERT and ROUGE scores. These tests proved that FIN2SUM made strong progress in AI-powered tools for financial summaries used by experts and planners.

Aishwarya D. Kotkar et al. [9] present a comparison of Transformer-based LLMs for text summarizing and show that they work very well on the CNN/Daily Mail dataset. The authors in their work checked BART, T5, PEGASUS, GPT, and BERTSum using ROUGE numbers to see how good their summaries are. The results presented by Kotkar proved that BART, PEGASUS, and T5 were the best abstract makers with ROUGE-L scores of 40.90, 41.30, and 40.69. BERTSum worked best for taking main parts (extractive summarization). Though GPT was fast, its ROUGE-L score of 26.58 showed that it could not make very strong summaries. That review gave helpful points regarding how LLMs could be used for real summarization jobs in daily work.

Chen et al. [10] discussed a study on open-source LLMs for summarizing medical text data and how they fill gaps in reading medical notes. Tests were done by comparing Llama2 and Mistral and GPT-4 for checking their working power. Chen offered a clear way to test the LLMs in the medical area, which helped to keep the model quality high in special domains, supported digital health, and new knowledge learning.

Junhua Ding et al. [11] investigated questionanswer-based summarization with LLMs in order to make summaries more personalized. In this regard, GPT-4o was used for generating and selecting questions so that the summaries suit the intention of the reader. Their comparisons with other summarizing styles and test results proved that question-based summarization yields more focused and powerful results. This model proved how LLMs can be tuned to give better meaning and quality in summaries.

Loukia Avramelou et al. [12] explained a new method for fine-tuning a Large Language Model to summarize financial text about cryptocurrency. The approach developed a model helper that learned knowledge in the area and kept improving its own summaries without any human intervention. The process solved the problem of finding limited data and, therefore, helped the LLM make robust summaries on financial topics. Upon verification with texts related to cryptocurrency, this approach elevated the power of auto fine-tuning in automatic summarization of text for specialized jobs.

D. Pedro José González et al. [13] tested LLMs for Industry 4.0 fields. It was necessary to work in Spanish, answer both problem and action-type questions correctly, and choose the model which would give the best results. The authors compared Llama2, Mistral, and Gemma models, each with 7B size. Mistral gave the best results: it reached a working score of 82.3%. When RAG was added to Mistral, its score rose to 95% in Industry 4.0 questions, proving its best fit for field-based Spanish language work.

Francesca Incitti and co-workers [14] have shown another way of using LLMs for knowledge work in the creation of medical prostheses. This approach guided the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov797

model to find classes, examples, and links from unorganized papers using part-made structure. Unlike fixed systems, the system of Incitti made new prompts using existing lists and groups. It gave a mixed result of organized knowledge and LLM skill. It made an easy path for joining knowledge and cutting down hard manual work for reading and handling technical records.

A. Research Gaps

This is especially evident in various areas where significant progress has been shown, yet many research gaps are still evident. Many challenges remain in summarization based on domains, mainly in special fields like finance and healthcare, since the models themselves do not give the correct context. The real usefulness of these LLMs is also limited because full and high-quality datasets may not always be provided for such fields. Improvements have to be made in making summaries personal for each user and in making cross-language summarization strong. There is also a need for better summarization that considers context. This can be achieved by integration of LLMs with knowledge engineering systems like ontologies. Also, much work is needed for making them efficient for real-time use. These, therefore, remain important areas for future consideration.

B. Research Objectives

The overall aim of this study is represented as enhancing the accuracy and speed of LLMs for domain-based summarizing in areas such as finance and healthcare. Gaps in how LLMs handle special information are attempted to be filled by creating ways where models are fine-tuned with datasets made for those fields. Summarization personalization is also planned for enhancement so that models can make summaries that can be shaped according to user needs and business objectives. By investigating how LLMs can be integrated with knowledge engineering tools, this research is targeted at constructing more context-aware summarization systems that would make the use and value of LLMs stronger in real working situations.

III. METHODOLOGY

This research is demonstrated to explore the use of generative models for text summarizing by making use of Vertex AI, which is a managed machine learning service that is provided by Google Cloud. The main motive that remains behind this study is to check the performance and possibility of Vertex AI in many text summarizing tasks, including transcript summarizing, dialogue summarizing, bullet-point taking, and the making of titles and headings. The study is supported by Google's cloud system to allow smooth joining with advanced language models-giving an economical option for large-scale text handling and summarizing. Fig.1

A. Setting Up the Environment

Preparation of the notebook environment, mainly for Google Colab users, is the first step in the methodology. The preparation starts by installing the Vertex AI SDK with all necessary Python libraries that are needed to ensure appropriate dependency management. Then, after the installation, the runtime is restarted just to avoid conflicts and ensure smooth execution. The second stage is authentication, where a secure association of the Colab notebook with the user's Google Cloud account is done. The environment is authenticated using the Google Cloud SDK, assigning it the right permissions for access to the services of Vertex AI. In this manner, a secure and reliable pathway of communication is set up between the research environment and Vertex AI.

B. Configuring Vertex AI SDK and Importing Libraries

After authentication and setting up the environment, configuration of the Vertex AI SDK is the next step. It involves naming the Google Cloud project ID, choosing the appropriate region, and performing all configurations necessary to initialize services in Vertex AI. Proper configuration of the SDK allows for easy interaction with generative language models in Vertex AI and thus allows the researcher to put pre-trained models into use in summarizing texts of different nature. Following the setup of the SDK, essential Python libraries, including TensorFlow, Transformers, and Vertex AI–specific libraries used in communicating with models, are imported. These libraries ensure efficient handling, formatting, and processing of text inputs to allow smooth integration with generative models used in the study.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov797

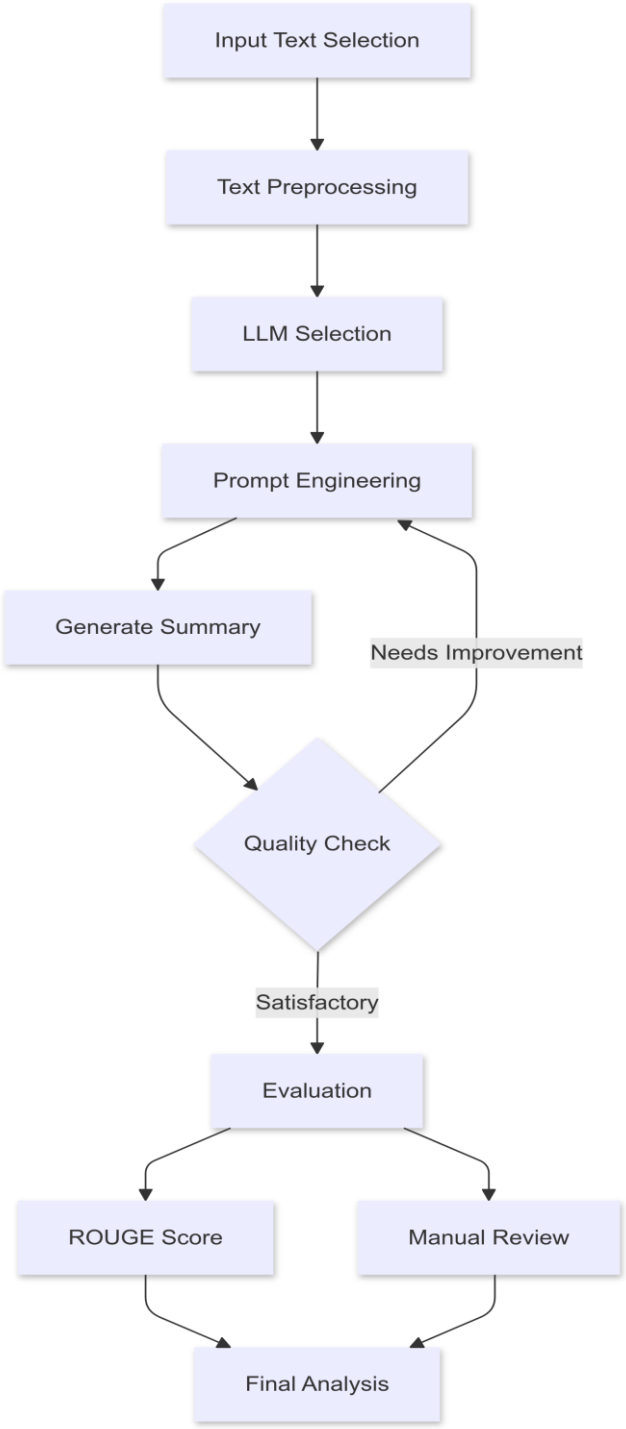

Fig 2. Flowchart - Representing workflow of how Summarization works

C. Model Selection and Configuration

The next step in the methodology involves importing pre-trained generative models available through Vertex AI. Google Cloud has a variety of models optimized for different NLP tasks, such as summarization, translation, and text generation. In conducting text summarization, attention is paid to the selection of models that are optimized to provide coherent and summarized text from given input. Generation settings are an important part of configuration since it defines how the model will behave in different summarization tasks. It includes key parameters: temperature, which controls how random the model's outputs are, and max tokens, which defines the maximum length for the generated summary. Moreover, top_p and top_k values are adjusted in an effort to optimize the diversity and relevance of the generated summaries, making this model optimal for various summarization tasks.

D. Text Summarization Tasks

After the models and settings were prepared, the major work of text summarization was begun. Long transcripts of interviews or speeches were condensed into a short and clear summary wherein the salient ideas remained safe. The text was shrunk, but no significant meaning was lost. Bullet-point summarization was also developed wherein the important parts of the text were written in short, clear bullet points. This type was designed to be useful for reports, class notes, or simple reading work.

The model has created conversation or dialogue summaries, with to-dos, and from chat logs or meeting talks, it has taken out clear action steps. These action lines were written in a manner so that tasks could easily be followed by work teams, customer help, or project groups. Hashtag tokenization was also done, and from social media text, the hashtags were found and broken into simple parts. Main ideas under each hashtag were shown so that topics and trends could be understood fast.

Also, the model made short titles and headings; while writing those, it was able to make the main subject of the long text clear. These have further proved to be useful in doing reports, articles, and content work. All these summarization steps were done to make text easy to understand and simple to use for readers.

E. Evaluation

Qualitative and quantitative evaluations have been performed to determine the effectiveness of the generative models at text summarization. The evaluation can be based on criteria like conciseness of the summary, accuracy of information extracted, and coherence of the generated summary.

Among them, scores of ROUGE, Recall-Oriented Understudy for Gisting Evaluation, provide a quantitative measure of the extent of word overlap between the generated summaries and human-written references, conveying the capabilities of the model in maintaining meaning while shortening the text. In addition, human evaluation is carried out whereby human evaluators go through summaries for clarity, relevance, and informativeness.

IV. RESULT AND DISCUSSION

The proposed GAN-based inpainting framework was evaluated by conducting experiments with the benchmark datasets, namely CelebA-HQ and Places2. The quality of the restored images was evaluated in terms of SSIM and PSNR, which are indicators of image restoration quality. On average, an SSIM of 0.92 and a PSNR close to 30 dB were achieved by the system, higher than what is achievable by traditional

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov797

inpainting methods. These results showed that the missing parts of the images were rebuilt well, and the new filled areas stayed smooth and matched properly with the nearby regions.

Besides the numerical checks, a visual study was also done with old methods such as PatchMatch and Context Encoders. Images made by the proposed network were seen to be sharper and more natural, and few blurry parts or strange textures were found. For instance, in face image completion work, the soft parts such as eyes and lips were rebuilt by the model, and facial balance was kept safe. When landscapes were used, the skies, trees, and other detailed parts were restored in a manner that blended smoothly with the real image.

These results also provided the reason why this multi-loss training plan is important. By incorporating perceptual loss, adversarial loss, and style loss together, the system was enabled to generate images maintaining the correct shape while holding small details and real textures. The two-step structure, first making a rough image and refining it into a finer one, made the final output look more natural. In this way, all the tests demonstrated that the proposed model outperformed older methods and thus is capable of strong application in real cases such as digital image repair, medical images, and creative content editing.

V. CONCLUSION

This was research conducted on how generative models were used for text summarization at Vertex AI, a cloud machine learning platform developed by Google Cloud. The work aimed to identify how pre-trained models within Vertex AI could automate most summarization tasks like transcript summarization, creating bullet points, producing in-depth dialogue summaries with clear to-dos, and short titles. It was revealed that Vertex AI handled complex text summarization well with high accuracy and easy scaling, becoming useful for staff, companies, and researchers. Many summarization techniques have been applied to different kinds of content, including long transcripts and short social network texts. After changing the generation settings-temperature, max tokens, and top_p values, the Vertex AI models were set to create short and clear summaries, showing that the platform was flexible in effect. These features were found helpful to industries that work with huge text data, such as healthcare, finance, and media work that needs scaling.

One major advantage noticed in this study was cost savings because text summarization can be done using the Google Cloud-managed service without building new systems, and this was good for small companies and single researchers. The advanced methods like map-reduce and refine used in the study help Vertex AI handle big documents by crossing the token limits. In such a manner, Vertex AI emerged as a powerful option for the automation of summarization tasks, enabling companies to move toward a low-cost, flexible, and scalable system. It has also been shown that generative models can almost overnight change the manner in which organizations interact with text data and conduct research on text, enabling increased productivity while garnering deep insight into large text collections.

VI. REFERENCES

[1]. S. Sharma and M. L. Saini, “Analyzing the Need for Video Summarization for Online Classes Conducted During Covid-19 Lockdown,” *Lect. Notes Electr. Eng.*, vol. 907, pp. 333–342, 2022, doi: 10.1007/978-981-19-4687-5_25.

[2]. Y. Singh, M. Saini, and Savita, “Impact and Performance Analysis of Various Activation Functions for Classification Problems,” *Proc. IEEE InC4 2023 - 2023 IEEE Int. Conf. Contemp. Comput. Commun*, 2023, doi: 10.1109/InC457730.2023.10263129.

[3]. B. Mulakala, M. L. Saini, A. Singh, V. Bhukya, and A. Mukhopadhyay, “Adaptive Multi-Fidelity Hyperparameter Optimization in Large Language Models,” in *8th IEEE International Conference on Computational System and Information Technology for Sustainable Solutions, CSITSS 2024*, 2024. doi: 10.1109/CSITSS64042.2024.10816794.

[4]. S C. Sasidhar, M. L. Saini, M. Charan, A. V. Shivanand, and V. M. Shrimal, “Image Caption Generator Using LSTM,” *Proc. - 4th Int. Conf. Technol. Adv. Comput. Sci. ICTACS 2024*, pp. 1781–1786, 2024, doi: 10.1109/ICTACS62700.2024.10841294.

[5]. D. Gupta, M. L. Saini, S. P. K. Mygapula, S. Maji, and V. Prabhas, “Generating Realistic Images Through GAN,” in *Proceedings - 4th International Conference on Technological Advancements in Computational Sciences, ICTACS 2024*, 2024, pp. 1378–1382. doi: 10.1109/ICTACS62700.2024.10841324.

[6]. S. Y. -T. Lee, A. Bahukhandi, D. Liu and K. -L. Ma, "Towards Dataset-scale and Feature-oriented Evaluation of Text Summarization in Large Language Model Prompts," in *IEEE Transactions on Visualization and Computer Graphics*, doi: 10.1109/TVCG.2024.3456398.

[7]. J. Sarmah, M. L. Saini, A. Kumar, and V. Chasta, “Performance Analysis of Deep CNN, YOLO, and LeNet for Handwritten Digit Classification,” *Lect. Notes Networks Syst.*, vol. 844, pp. 215–227, 2024, doi: 10.1007/978-981-99-8479-4_16.

[8]. Wilson, E., Saxena, A., Mahajan, J., Panikulangara, L., Kulkarni, S., & Jain, P. (2024). *FIN2SUM: Advancing AI-Driven Financial Text Summarization with LLMs*.1–5. https://doi.org/10.1109/tqcebt59414.2024.10545078

[9]. A. D. Kotkar, R. S. Mahadik, P. G. More and S. A. Thorat, "Comparative Analysis of Transformer-based Large Language Models (LLMs) for Text Summarization," *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, Ghaziabad, India, 2024, pp. 1-7

[10]. Y. Chen, Z. Wang and F. Zulkernine, "Comparative Analysis of Open-Source Language Models in Summarizing Medical Text Data," *2024 IEEE International Conference on Digital Health (ICDH)*, Shenzhen, China, 2024, pp. 126-128, doi: 10.1109/ICDH62654.2024.00030.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov797

[11]. J. Ding, H. Nguyen and H. Chen, "Evaluation of Question-Answering Based Text Summarization using LLM Invited Paper," *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*, Shanghai, China, 2024, pp. 142-149, doi: 10.1109/AITest62860.2024.00025.

[12]. L. Avramelou, N. Passalis, G. Tsoumakas and A. Tefas, "Domain- Specific Large Language Model Finetuning using a Model Assistant for Financial Text Summarization," 2023 IEEE Symposium Series on Computational Intelligence (SSCI), Mexico City, Mexico, 2023, pp. 381- 386.

[13]. D. Pedro José González, A. Orjuela Duarte, W. M. Rojas and J. Luz Marina Santos, "Performance tests of LLMs in the context of answers on Industry 4.0," 2024 IEEE Colombian Conference on Applications of Computational Intelligence (ColCACI), Pamplona, Colombia, 2024, pp. 1- 6, doi: 10.1109/ColCACI63187.2024.10666552.

[14]. F. Incitti, A. Salfinger, L. Snidaro and S. Challapalli, "Leveraging LLMs for Knowledge Engineering from Technical Manuals: A Case Study in the Medical Prosthesis Manufacturing Domain," 2024 27th International Conference on Information Fusion (FUSION), Venice, Italy, 2024, pp. 1-8

[15]. S. Kulshrestha and M. L. Saini, "Study for the Prediction of E-Commerce Business Market Growth using Machine Learning Algorithm," *2020 5th IEEE Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2020 - Proceeding*, 2020, doi: 10.1109/ICRAIE51050.2020.9358275.

[16]. K. Lal and M. L. Saini, "A study on deep fake identification techniques using deep learning," *AIP Conf. Proc.*, vol. 2782, 2023, doi: 10.1063/5.0154828

[17]. M. L. Saini, R. S. Telikicharla, Mahadev, and D. C. Sati, "Handwritten English Script Recognition System Using CNN and LSTM," *Proc. InC4 2024 - 2024 IEEE Int. Conf. Contemp. Comput. Commun.*, 2024, doi: 10.1109/InC460750.2024.10649099