AI Enhanced Image to Audio Encryption

Anjali Chennupati, Bhamidipati Prahas, Bharadwaj Aaditya Ghali, Dr. Nidhin Prabhakar. T. V
Department of Computer Science and Engineering,
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India

anjali.m.chennupati@gmail.com, prahasb154@gmail.com, ba.aaditya03@gmail.com, tv_nidhin@blr.amrita.edu

***Abstract*— This innovative project addresses the critical need for secure multimedia data sharing by integrating image-to-audio encryption with advanced AI-based data-hiding techniques. The core concept involves transforming static visual content into dynamic auditory experiences, injecting semantic depth and artistic creativity into the resulting audio output. Unlike conventional methods, this approach goes beyond translating pixels into sound waves. It holds promise in multimedia communication, artistic expression, and data protection. To enhance security, the project introduces AI-enhanced data-hiding techniques, leveraging adversarial training and reinforcement learning. This extra layer ensures resistance to unauthorized access and tampering. In summary, this project fuses image-to-audio encryption and AI-enhanced data hiding, aiming to revolutionize multimedia. It offers a robust solution, transcending boundaries by safeguarding visual data through the immersive medium of sound.**

Keywords—Image to audio encryption and decryption, Neural Networks, Steganography, hiding.

I. INTRODUCTION

In an era marked by the exponential growth of multimedia data sharing and the paramount importance of data security, the fusion of image-to-audio encryption with cutting-edge artificial intelligence (AI) techniques emerges as a pioneering solution. This innovative approach transcends traditional methods by seamlessly integrating advanced AI technology into the encryption process, elevating security and robustness to unprecedented levels. At its core, image-to-audio encryption represents a groundbreaking paradigm shift, transforming static visual content into dynamic auditory experiences. However, this approach goes beyond mere translation of pixels into sound waves; it imbues semantic relevance and artistic depth into the auditory output, enhancing its value and impact.

Building upon the foundation of image-to-audio encryption, this research endeavours to introduce AI-enhanced data-hiding techniques. Leveraging state-of-the-art AI methodologies, such as Adaptive Steganography where AI is employed to dynamically adjust data embedding strength based on the audio's characteristics, ensuring that the hidden data remains concealed while adapting to different audio types. This fusion of image-to-audio encryption with AI-driven data hiding not only advances the realm of multimedia security but also promises diverse applications, from multimedia communication to artistic expression and comprehensive data protection. This introduction sets the stage for an exploration of the transformative potential of this innovative approach.

In the context of comprehensive data protection, the project can emphasize its contribution to establishing a multi-layered security paradigm. The combination of image-to-audio encryption and AI-enhanced data hiding creates a sophisticated defence mechanism against unauthorized access and tampering. The adaptability of the system to different audio characteristics ensures a versatile and resilient solution that transcends traditional encryption methods.

As a forward-looking statement, the research could touch upon the scalability and potential future applications of the developed framework. The ever-evolving landscape of AI and multimedia technologies suggests that this innovative approach could find relevance in emerging fields such as augmented reality, virtual communication, and interactive digital arts, contributing to the continuous evolution of secure and creative data handling.

Section 2 presents strategies and methodologies explored for the image-to-audio encryption that are noteworthy. Section 3 describes the proposed methodology for the steganography process and image-to-audio encryption and decryption, Section 4 presents the results and analysis and finally, Section 5 concludes the findings and provides pointers towards future directions.

II. RELATED WORKS

Image-to-audio conversion holds significant importance in the realm of multimedia technology and communication. This innovative process allows for the transformation of visual information into auditory experiences, introducing a novel dimension to data representation and communication. The significance lies in its diverse applications, ranging from accessibility features for visually impaired individuals to creative artistic expression. By converting images into sound, this technique facilitates a unique form of data representation that goes beyond traditional visual mediums. Additionally, in the context of secure communication, image-to-audio conversion can serve as a discreet method of data encryption and steganography, where sensitive information is hidden within audio files. This not only adds a layer of security to digital communication but also opens up avenues for creative and secure multimedia applications, contributing to the evolving landscape of information technology.

Deep learning techniques are advancing secure communication and identity verification by embedding information in noise-tolerant signals like audio, video, and images. Zihan et al. [1] performed a survey that covers deep learning techniques for data hiding, including watermarking and steganography, emphasizing their common goals and potential benefits of deep learning. It explores future research directions to enhance digital IP protection and communication security in Responsible AI. Deep learning is expected to revolutionize digital security, surpassing traditional methods and enhancing accountability and safety in AI.

Savita et al. [2] present an asymmetric audio and image encryption method using QR decomposition and random modulus decomposition in the Fresnel domain. Audio is

transformed into a sound map, which is then encrypted. The scheme is validated for multiple audios and grayscale images, demonstrating robustness and sensitivity to input parameters against various attacks. The research highlights the scheme's strength through noise and attack analysis and positions it as an efficient method for audio encryption, encouraging further exploration in this interdisciplinary field.

Text-conditioned generative models are another method to generate images from audio descriptions, producing high-quality images that is proposed by Yaris et al. [3]. An audio encoding model is introduced to bridge audio and text representations with few trainable parameters, outperforming baseline methods in objective and subjective metrics for image generation. An initial step is marked in audio-conditioned image generation, emphasizing the richness of hidden audio information, suggesting the need for more community attention to this problem.

Another method that is used in image-to-audio conversion is steganography which involves concealing information within audio files derived from images. Subhajit et al. [4] focus on concealing images within audio, converting audio steganography into image steganography using mel-spectrograms as the cover medium. Deep neural networks (DNNs) are brought in for this purpose, optimizing perceptual quality. The proposed technique effectively hides images within audio, evades steganalysis tools, handles various image types, and allows multiple image data to be concealed in a single audio file. Image compression and generative models are employed, effectively hiding and reconstructing messages while being robust to various attacks. Future work may extend the model to video steganography and enhance resistance to deep neural network-based steganalysis.

Dalal et al. [5] present a novel approach for embedding image data into audio files using LSB-based Audio Steganography, with encrypted image placement based on the image's encrypted bytes in the audio's low-frequency range. The method demonstrates improved results through subjective and objective testing, achieving a high Peak Signal-to-Noise Ratio (PSNR) of 61.863 and a low Mean Squared Error (MSE) of 0.329. The research successfully embeds and extracts image data within audio, focusing on audio steganography. This technique enhances security, making it challenging for intruders to detect the hidden image within the audio.

An alternate technique that is employed in image-to-audio conversion is Reversible Data Hiding (RDH) which is a technique that allows for the embedding of additional data within audio signals derived from images while maintaining the ability to perfectly recover the original image. Zhongyun et al. [6] suggest a reversible data hiding in encrypted images (RDH-EI) scheme with multiple data hiders to enhance security. A matrix-based secret sharing technique is utilized to create an (r, n)-threshold RDH-EI scheme, where the content owner encrypts an image into n encrypted images and distributes them to data hiders. The scheme provides robustness against n-r points of failure, ensuring image content confidentiality and increased embedding capacity.

Chunqiang et al. [7] present another novel Reversible Data Hiding in Encrypted Images (RDHEI) method with hierarchical embedding for high payload. A hierarchical label map generation technique is established that divides prediction errors into small, medium, and large-magnitude errors. Experimental results indicate that this method outperforms state-of-the-art RDHEI techniques with average payloads of 3.4568 bpp and 3.6823 bpp on BOWS-2 and BOSSbase datasets, respectively.

Roy et al. [8] explore an innovation in this domain. The proposed IM2WAV system utilizes two Transformer language models and a VQ-VAE-based model to generate semantically relevant audio from input images. A CLIP model is leveraged for visual representation conditioning and employs a classifier-free guidance method to enhance generation alignment with the input image. IM2WAV outperforms baselines in fidelity and relevance metrics, as demonstrated through comprehensive evaluation and an ablation study, while introducing IMAGEHEAR, an out-of-domain image dataset, as a benchmark for future image-to-audio models.

Security concerns in image-to-audio encryption are paramount, given the potential vulnerabilities that may arise during the process of transforming visual information into auditory representations. One significant challenge is ensuring the robustness of encryption algorithms to prevent unauthorized access or tampering. Arslan et al. [9] address this very issue. A method in proposed that incorporates chaos for image scrambling and uses Discrete Wavelet Transform (DWT) to encrypt low-frequency bands efficiently. Performance tests including entropy, correlation, and noise resistance indicate the algorithm's strong security, but it may allow partial content visualization in decrypted images.

A breakthrough application of image-to-audio encryption is explained by Hemalatha et al. [10] by presenting a novel Blind Book reader system for visually challenged individuals, offering a solution to read books and newspapers. A method is established that utilizes a webcam to capture images, convert them into text, and then into voice, which can be translated into regional languages. The system, based on OCR algorithms, demonstrates superior accuracy compared to existing methods.

The steganography inspiration chosen for this project is partially from the procedure proposed by Shumeet Baluja [11]. The study investigates hiding big colour pictures inside other, same-size ones. For this, it uses deep learning networks to do the job quietly without being noticed. Unlike old ways, it spreads out how the secret picture is viewed across all available parts. This makes effectiveness better on different natural photos from places outside ImageNet.

In conclusion, image into audio steganography as explored by W. Cui et al. [14] offers a unique and versatile approach to secure communication and data protection. Its potential applications range from confidential information exchange to steganography, but careful consideration of robust encryption methods and key management is crucial to address security concerns. As technology evolves, continual advancements in image-to-audio encryption techniques will play a vital role in ensuring the confidentiality and integrity of multimedia data in the digital landscape.

III. PROPOSED METHODOLOGY

The system depicted in Fig 1. serves as the foundational framework for the AI-enhanced image-to-audio encryption proposed in this research. The proposed methodology begins with steganography, the output of which is given as input for the audio encryptor. The audio encryptor takes the input image and samples it into bytes and an audio of various lengths is

generated from these bytes. The audio decryption involves converting the audio into byte samples and then reshaping it to the desired image size. The decrypted output is given as input to the steganography model so that the secret image is extracted. This proposed system mainly comprises of two sub-systems i.e. image in image steganography and image-to-audio encryption. In this context, steganography refers to the practice of concealing the input image known as the secret image, within another image known as the cover image in such a way that the alteration is not known to the human eye. The process involved is inserting the content of the secret image into the cover image, producing a steganographic image. The main aim is to make the steganographic image indistinguishable from the original cover image. Now delving into the first sub-system, it primarily consists of two networks that utilize the power of neural networks, the Hide-Net and the Reveal-Net.

The Hide-Net is an Unet-based Generator, in image-to-image translation, concealing information within images through steganography, which is characterized by its U-shaped convolutional neural network architecture that allows the preservation of high-resolution details during up-sampling. The structure of the proposed Unet model comprises of two main classes. The first is the UnetGenerator class. In the UnetGenerator Class, the generator begins with the innermost block and iteratively adds skip connection blocks with the help of the second class, UnetSkipConnectionBlock. Each block consists of various layers such as down-sampling, up-sampling, normalization, and activation layers. The outermost block includes the output activation function in order to generate the final output. The forward pass function involves sequentially applying the defined blocks to the input tensor.

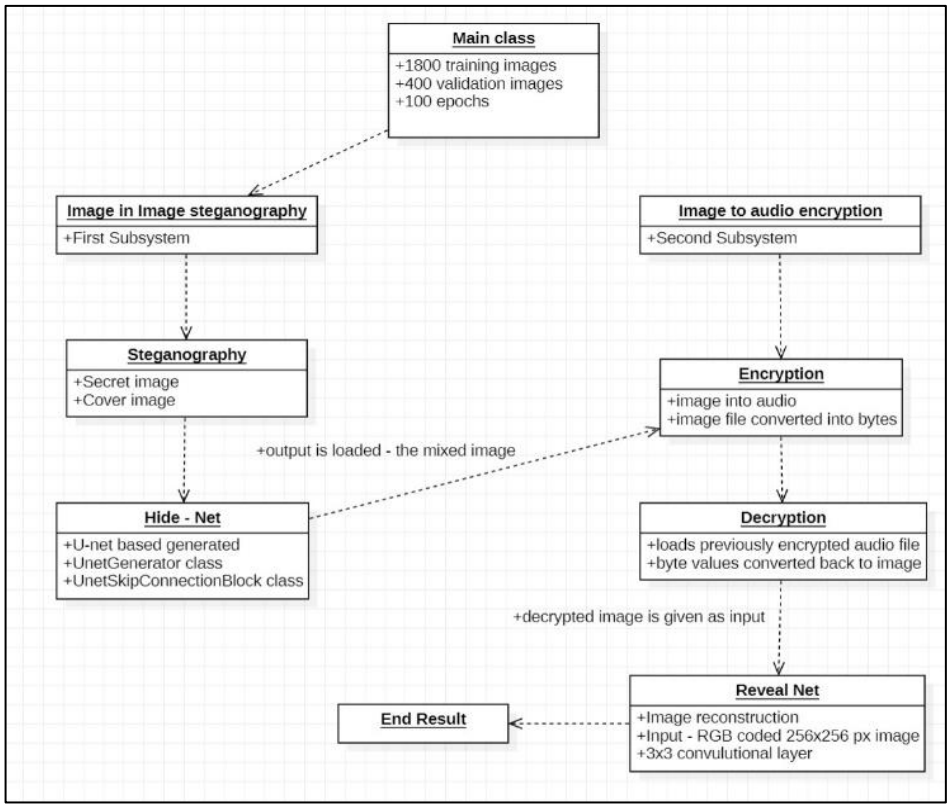

Fig 1. Block diagram for image-to-audio encryption and decryption

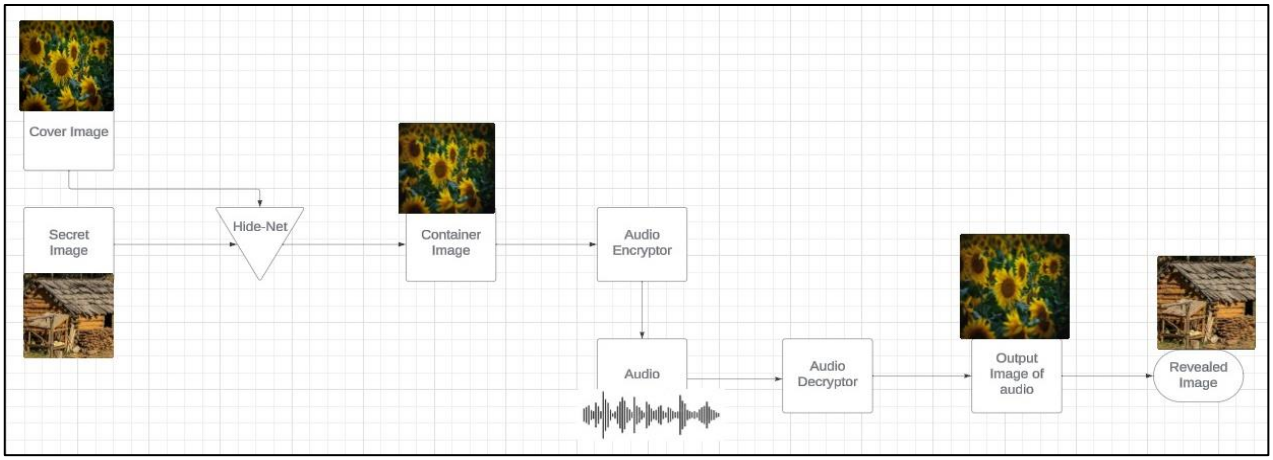

Fig 2. The encryption and decryption process in detail

The output is produced using the U-shaped architecture with skip connections. The UnetSkipConnectionBlock class is a submodule of Unet. All the blocks comprise a normalization for the inner layer, an up-sampling convolutional layer, LeakyReLU activation, and a down-sampling convolutional layer. Depending on whether a block is intermediate, outermost, or innermost, its configuration changes. The preset layers are applied one after the other in the forward pass function. It concatenates the input with the processed tensor for the outermost and intermediate blocks, omitting the connections.

The second segment of the first subsystem is Reveal–Net, which is designed with the primary objective of performing image generation or reconstruction. It is structured to transform an input tensor into an output tensor representing the revealed image. The network comprises of a series of convolutional layers, batch normalization, and Rectified Linear Unit (ReLU) activation function. The input in this phase must be an RGB-coded image of 256x256 pixel dimension. A 3x3 convolutional layer with the number of channels (nc = 3) and number of filters (default nhf = 64) is where the network starts. The number of filters is gradually reduced and then doubled in subsequent layers, creating a smooth transition between encoding and decoding. After every convolutional layer, batch normalization layers are employed to enhance the training process's stability and convergence. Similar blocks follow gradually changing the number of filters in the subsequent layers. The Rectified Linear Unit (ReLU) activation function adds non-linearity to the network. The Sigmoid activation function is the default configuration for the network's last layer.

These networks are trained and validated using 1800 training images and 400 validation images from the Linnaeus 5 image dataset. The training is done for 100 epochs while calculating the loss of the hide-net and the hide of the reconstruction loss. The epoch 99 loss of Hide-Net is netH_epoch_99, sumloss = 0.001365, h-loss = 0.000897, and the epoch 99 loss of Reveal-Net is netR_epoch_99, sumloss = 0.001365, Rloss = 0.000623. The h-loss value is used to compute the error between the network's predictions and the actual value. The goal of training a neural network is to reduce the h-loss and r-loss value. Fig 3 and Fig 4. show that the h-loss and r-loss value decreases as the number of epochs increases. This shows that the neural network can learn from the data and improve its predictions.

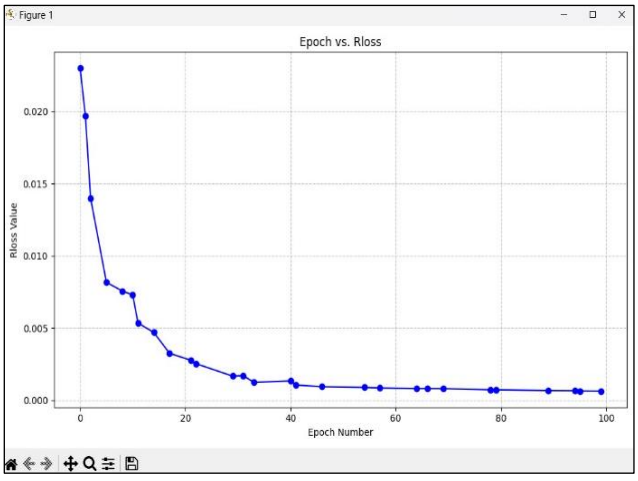

Fig 3. Epoch vs r-loss

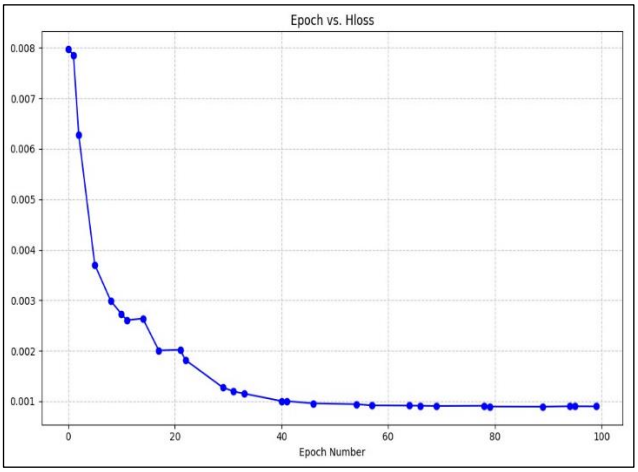

Fig 4. Epoch vs h-loss

Fig 5. shows the input cover image and the secret image on the top left and right respectively. The second row consists of the container image obtained after epoch 10, which comprises of the steganography image and the revealed image on the left and right respectively. The inference drawn from the above is that the h-loss and Rloss have decreased. As depicted in Fig 6, the hiding and reconstruction of the image has improved significantly, from epoch 10. It can be inferred from Fig 5. and Fig 6., that a lesser number of epochs does not result in the expected output as opposed to a much larger number of epochs which results in the desired output.

Fig 5. Epoch 10 of the training process.

Fig 5. Epoch 99 of the training process.

Transitioning into the second subsystem, it is used for image-to-audio encryption and decryption. The structure of this system consists of two parts the encryption and the decryption. The encryption process starts by converting the

input image into a bmp image file and this bmp file is then loaded. This image file is converted into bytes format and normalized between the range -1.0 to 1.0. It is then scaled to a 16-bit signed integer. The audio segment is created using the normalized and scaled image bytes and is saved as a .wav file. The decryption process starts by loading the previously encrypted audio file. Audio samples are converted to 16-bit samples and byte values. These byte values are used to convert back to an image using the PIL library. This decrypted image is saved as a .png file.

In conclusion, the first subsystem does image-in-image steganography using Hide Network. The output of the Hide-Net's steganographic image is fed into the second subsystem that does image-to-audio encryption and decryption. The image is encrypted into an audio file. This audio file is given as input to the decryption process. After decryption the decrypted image output is used as input by the Reveal Net and the secret image is decrypted. The two subsystems when combined provide a model for AI-enhanced image steganography with audio encryption. Through the fusion of cutting-edge techniques, it not only transforms visual content into auditory experiences but also promises versatile applications in communication, artistic expression, and robust data protection. The model's success is demonstrated in its ability to combine several models in a single forum. Such a specialty helps multimedia privacy and security to become more easily achieved.

IV. RESULTS AND ANALYSIS

The described system employs a framework, leveraging the power of steganography and image-to-audio encryption and decryption. The Unet (Hide-Net) and Reveal Net (Rnet) processes play pivotal roles in the AI-enhanced image-to-audio encryption process. The Unet is a U-convolutional neural network and works with skip connection to maintain the fine details from the high-resolution algorithm if they've remained unaltered until reaching low resolutions. It serves as the least significant bit hiding network, combining both cover image and hidden content together to create a new carrier picture.

On the other hand, the Reveal Net reconstructs the hidden image from the container, employing convolutional layers and activation functions. Together, these processes form a robust system for concealing and revealing information within multimedia data through advanced neural network architectures. After the concealing and reconstruction of the image, the next phase of image-to-audio encryption begins. The output image post steganography is converted into 16-bit signed integer and later mapped into an audio. This audio is again decrypted to retrieve the concealed image.

As a part of this comprehensive methodology, it is crucial to address the clarity of the output image produced after the encryption and decryption process. The initial comparison can be done between the cover image, i.e., the image that the secret image is overlapped with, and the container image, i.e., the output image produced after the Hide–Net process as depicted in Fig 4. StegExpose, a steganalysis toolkit, was used to test the detectability of the image. The steganography performed in the container image was undetected when it was passed through this toolkit due to the fact that the absolute difference was close to null with a loss value of 0.0008769232663325965.

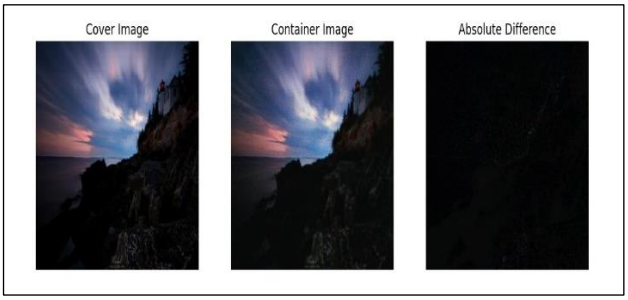

Fig 4. Absolute difference between cover image and output of Hide-Net

The second phase of comparison can be performed between the secret image and the revealed image that is obtained post the Reveal-Net process as depicted in Fig 5. The absolute difference between the images was close to null as well with a loss value of 0.0026543575804680586.

In summary, the evaluation of the image-to-audio encryption and decryption processes reveals a highly effective concealment and retrieval of information. This robustness in both encryption and decryption processes establishes the reliability and security of the proposed image-to-audio framework.

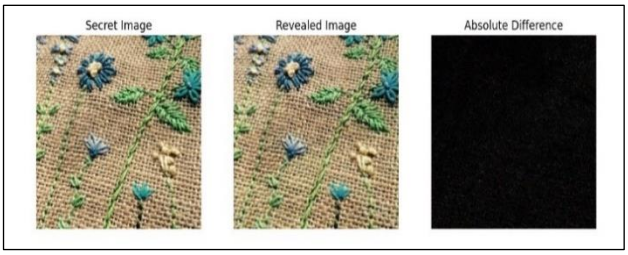

Fig 4. Absolute difference between secret image and output of Reveal - Net

CONCLUSION AND FUTURE

In conclusion, this innovative project combines image-to-audio encryption with progressive AI methods. It creates a strong framework for protected media messages and calls. The Hide-Net (Unet) and Reveal Net parts show how strong neural network structures are such that it can hide or expose information in multimedia data. This new formula goes beyond old methods of hiding messages, changing still pictures into sounds you can hear. This improves safety and makes things more creative at the same time.

Looking forward, the future potential of this project is progressive. Making the neural network better and faster, looking at new structures, adding in smart AI techniques could improve how well this system works. Also, progress can be made by making it work for many different types of media and real-time handling to make it more useful in everyday life. Using smart hiding methods in steganography, where AI changes how data is hidden based on audio features, can strengthen security. Working with experts in code and audio can give important information to make the security process stronger. In addition, creating easy-to-use screens and making them work with different systems could help more people use it. As keeping security in multimedia becomes very important, this project starts the effort to do more research and improvements on secure systems that use AI for communication.

REFERENCES

[1] Wang, Zihan, et al. “Data Hiding With Deep Learning: A Survey Unifying Digital Watermarking and Steganography.” IEEE Transactions on Computational Social Systems 10 (2021): 2985-2999.

[2] SAVITA ANJANA1*, AK YADAV2, PHOOL SINGH2, HUKUM SINGH “Audio and image encryption scheme based on QR decomposition and random modulus decomposition in Fresnel domain.” 123031- Optica Applicata, Vol. LII, No. 3, 2022

[3] Yariv, Guy, Itai Gat, Lior Wolf, Yossi Adi, and Idan Schwartz. 2023. “AudioToken: Adaptation of Text-Conditioned Diffusion Models for Audio-To-Image Generation.” ArXiv.org. May 22, 2023. https://doi.org/10.48550/arXiv.2305.13050.

[4] Paul, S., Mishra, D. Hiding images within audio using deep generative model. *Multimed Tools Appl* **82**, 5049–5072 (2023). https://doi.org/10.1007/s11042-022-13034-4

[5] Hmood, Dalal & Abbas, Khamael & Altaei, Mohammed. (2012). A New Steganographic Method for Embedded Image In Audio File. International Journal of Computer Science and Security. 6. 135.

[6] Z. Hua, et al.,"Matrix-Based Secret Sharing for Reversible Data Hiding in Encrypted Images" in IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 05, pp. 3669-3686, 2023.

[7] C. Yu, X. Zhang, X. Zhang, G. Li and Z. Tang, "Reversible Data Hiding With Hierarchical Embedding for Encrypted Images," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 2, pp. 451-466, Feb. 2022, doi: 10.1109/TCSVT.2021.3062947.

[8] Sheffer, Roy & Adi, Yossi. (2022). I Hear Your True Colors: Image Guided Audio Generation. 10.48550/arXiv.2211.03089.

[9] A. Shafique, M. M. Hazzazi, A. R. Alharbi and I. Hussain, "Integration of Spatial and Frequency Domain Encryption for Digital Images," in IEEE Access, vol. 9, pp. 149943-149954, 2021, doi: 10.1109/ACCESS.2021.3125961.

[10] Hemalatha, B., Karthik, B., Balaji, S., Vijayalakshmi, G., Shaw, R.N. (2022). A Novel Approach for Blind - Image to Audio Conversion in Regional Language. In: Mekhilef, S., Shaw, R.N., Siano, P. (eds) Innovations in Electrical and Electronic Engineering. ICEEE 2022.

[11] Baluja, Shumeet. “Hiding Images in Plain Sight: Deep Steganography.” Neural Information Processing Systems (2017).

[12] Nidhin PTV, Hemanth VK, Sachin Kumar S, KP Soman, Arun S, Comparative Study of Recent Compressed Sensing Methodologies in Astronomical Images, Eco-friendly Computing and Communication Systems Communication in Computer and Information Science (Springer), Vol 305, pp 108-116, 2012 (Scopus).

[13] Midhun EM, Sarath R Nair, Nidhin PTV, Sachin Kumar S, Deep Model for Hyperspectral Image Classification using Restricted Boltzman Machines, International Conference on Interdisciplinary Advances in Applied Computing, ACM-Sig Conference, 2014 (Scopus)

[14] W. Cui, S. Liu, F. Jiang, Y. Liu and D. Zhao, "Multi-Stage Residual Hiding for Image-Into-Audio Steganography," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 2832-2836, doi: 10.1109/ICASSP40776.2020.9054033.

[15] Deep model for classification of hyperspectral image using restricted boltzmann machine, M. E. Midhun, Sarath R Nair, V. T. Nidhin Prabhakar, S. Sachin Kumar, 10 October 2014.

[16] A. K. Sahu and M. Sahu, "Digital Image Steganography and Steganalysis: A Journey of the Past Three Decades," September, Open Computer Science, Degruyter, 10, 1-47, DOI: https://doi.org/10.1515/comp-2020-0136, 2020.

[17] P. Mathivanan and A. B. Ganesh, "Colour image steganography using XOR multi-bit embedding process," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 1980-1988, doi: 10.1109/ICECDS.2017.8389797.

[18] Kavya Duvvuri, P Nandieswar Reddy, Harshitha Kanisettypalli, Radha D, Nidhin Prabhakar T V "Comparative Analysis of Pattern Matching Algorithms Using DNA Sequences", 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon).