Received August 6, 2021, accepted August 24, 2021, date of publication August 26, 2021, date of current version September 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108183

A Novel Approach for Linguistic Steganography Evaluation Based on Artificial Neural Networks

R. GURUNATH[1], AHMED H. ALAHMADI[2], DEBABRATA SAMANTA[1], (Member, IEEE), MOHAMMAD ZUBAIR KHAN[2], AND ABDULRAHMAN ALAHMADI[2]

[1]Department of Computer Science, CHRIST (Deemed to be University), Bengaluru, Karnataka 560029, India
[2]Department of Computer Science and Information, Taibah University, Medina 42353, Saudi Arabia

Corresponding authors: Debabrata Samanta (debabrata.samanta369@gmail.com) and Mohammad Zubair Khan (mkhanb@taibahu.edu.sa)

ABSTRACT Increasing prevalence and simplicity of using Artificial Intelligence (AI) techniques, Steganography is shifting from conventional model building to AI model building. AI enables computers to learn from their mistakes, adapt to emerging inputs, and carry out human-like activities. Traditional Linguistic Steganographic approaches lack automation, analysis of Cover text and hidden text volume and accuracy. A formal methodology is used in only a few Steganographic approaches. In the vast majority of situations, traditional approaches fail to survive third-party vulnerability. This study looks at evaluation of an AI-based statistical language model for text Steganography. Since the advent of Natural Language Processing (NLP) into the research field, linguistic Steganography has superseded other types of Steganography. This paper proposes the positive aspects of NLP-based Markov chain model for an auto-generative cover text. The embedding rate, volume, and other attributes of Recurrent Neural Networks (RNN) Steganographic schemes are contrasted in this article between RNN-Stega and RNN-generated Lyrics, two RNN methods. Here the RNN model follows Long Short Term Memory (LSTM) neural network. The paper also includes a case study on Artificial Intelligence and Information Security, which discusses history, applications, AI challenges, and how AI can help with security threats and vulnerabilities. The final portion is dedicated to the study's shortcomings, which may be the subject of future research.

INDEX TERMS Artificial intelligence, steganography, linguistic steganography, statistical language model, natural language processing, NLP, Markov chain model, recurrent neural networks, RNN, LSTM.

I. INTRODUCTION

Steganography is the practice of sending messages or information to a receiver that is hidden in an entity that may take the form of a letter, a photograph, document, or a variety of other types. Steganography is derived from two Greek words: Stegos, which means "to cover," and grayfia, which means "books," which means "secret writing." When uploading data over the Internet, Steganography is one of the most popular ways to enhance security and data safety. The primary goal of Steganography is to hide messages in communication media using text, graphic, audio and video [7], [36]. As seen in Figure 1, carriers' files may be categorised as text, image, and audio, video, or protocol files.

Although textual documents have less redundant data than other digital media including images, audio, or video files, the most challenging type of Steganography is text Steganography. As seen in the Figure 2, Format-based approaches, random and statistical generation, and linguistic approaches are the three types of text Steganography [26] [4].

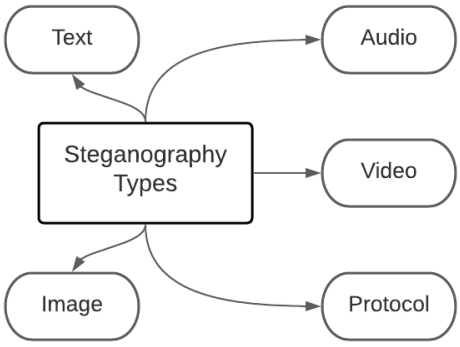

FIGURE 1. Digital steganography types.

The use of physical text encoding to mask information is used in format-based approaches. Random and statistical generation extracts a language's statistical properties, which

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

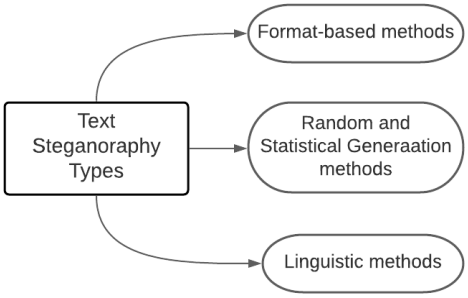

FIGURE 2. Types of text steganography.

are then used to produce cover documents. Text Steganography takes a linguistic view, which considers the use of language properties in text alteration. The two forms of linguistic Steganography (see Figure 3) are carrier-modification based Steganography (CMS) and carrier-generation based Steganography (CGS). To embed special secret information, the CMS strategy mainly consists of replacing lexical or sentence-level semantic units in the text with synonyms. Based on the secret knowledge that has to be sent, the CGS approach will swiftly produce a semantically rich and normal-looking carrier. When it comes to concealing sensitive information in the carrier, this feature provides additional options, allowing for a better level of information concealment capabilities [39]. For digital photos, a new transform domain steganography approach based on integer wavelet transform (IWT) was utilised, as well as a chaotic map. This map is a modified logistic map that improves the suggested method's key length and security. The proposed method has a strong potential of hiding information in any photos used as a cover media, according to experimental results [32], [33].

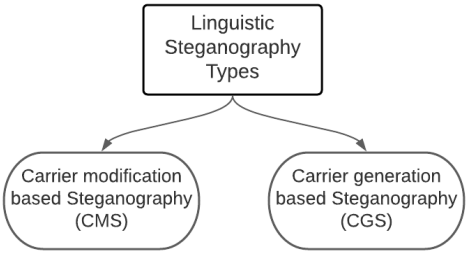

FIGURE 3. Linguistic steganography types.

In recent years, researchers have been increasingly interested in the text data hiding approach based on cover-less text Steganography. It could conceal sensitive information by producing new text that adheres to natural language's statistical characteristics. The Markov chain model is one such example of passing hidden information using the model and later producing Steganographic sentences. A discrete stochastic process is a Markov chain. It denotes a set of states in which, based on present facts or understanding, the prior historical condition of the mechanism cannot be used to forecast the future state. The mechanism will switch from one state to another or stay in the current state at each stage of the Markov chain, depending on the probability distribution [34].

The Markov model approximates the language model and calculates every word's conditional probability distribution. Yet, owing to the lacunas of the Markov model, the consistency of the text provided by the Markov model [38] is precisely insufficient, making it easy to recognise. With the advancement of NLP technologies in late years, an incremental Steganographic text generation models based on neural network models have appeared. A neural network may be replaced with a Markov model as a process progression [32], [33]. To do so, fragment the dictionary and precisely translate each word, then learn the statistical language model of natural text using the recurrent neural network (RNN). At the end, various words are chosen as output at each stage of the automated text generation process, depending on the information that needs to be concealed. However, it seems that this is insufficient. Tests recently demonstrated that, since current models fail to monitor the exact meaning of the produced Steganographic texts, though the produced text is accurate enough, it still poses a security danger. Analyzing the technologies of automated Steganographic text creation with controllable semantics is a problem which must be addressed. The sections in this article follow are Literature Review, Artificial Intelligence and Information Security: A Case Study, Natural Language processing in Steganography- Markov model, and Comparison of Two RNN Steganography Methods.

II. MOTIVATION

The goal of this research is to find solutions to problems of existing data concealing techniques' basic characteristics (undetectability and payload capacity). Modern data concealment places a greater emphasis on automatically produced cover text rather than existing cover texts. In this study, we looked at how Artificial Intelligence approaches like RNN, LSTM, Markov chain model, and Huffman tree may be used to generate automated text that is superior to traditional methods.

III. CONTRIBUTION

The main contribution of our research to find the research gaps identify in the current study with better solution.

IV. LITERATURE REVIEW

Chang and Clark [6] critically says that, as the usage of the internet and smart phones expands, Social media produces a vast amount of data every day. As a result, transmitting private information via social media apps attracts less attention, making this method of contact more reliable. Steganography methods' efficacy is dependent on the non-detection of hidden material by third parties. Otherwise, it seems that the techniques are ineffective. The increasing popularity and

scale of text data is fuelled by the central presence of Internet information in human livelihoods as well as the growth of text-based information exchange platforms like email, forums, and Facebook messenger and other social media applications. This rise in the importance of electronic text, in particular, raises questions regarding the use of text media as a hidden contact medium.

Xiang *et al.* [35] Finds that, Languages based on linguistics that were used in the early stages of text production Steganography made use of template creation based on rules, such as reference embraced context-free grammar. Natural language processing technology is used in recent text generation Steganography innovations; many models are available, with the Markov model being one of them. Poor embedding capability can be an issue for most modern linguistic Steganographic methods. Linguistic Steganography based on text creation, on the other hand, efficiently overcomes this difficulty. This approach would not necessitate the development of an initial text in preparation. Such a system has a high concealing ability as there is more provision for storing confidential information and there is no maximum bound to the size of the created text. Conventional text generation approaches lacked the artificial intelligence needed to produce greater text automatically. The Stego text that resulted was prone to mistakes that challenged logic and common sense.

Wu *et al.* [34] opinion that, deep learning-based text generation models have been brought into the area of linguistic steganography by researchers. The Markov chain model was used to accurately calculate the frequency of each term in the training set. The Markov chain dependent steganography model can be used to effectively deal with third-party attacks based on statistical properties of natural language, as well as to produce the best possible sentences. The deep learning touch can vastly increase the consistency and embedding quality of the Stego text. However, it uses more energy and is slower to run.

Grosvald and Orgun [14] strongly says, cover text systems have a few weaknesses. To begin with, if the message's cover text is a well-known text, someone intercepting it would be aware of it. The fact that the cover text seems to be identical but isn't poses certain concerns right away. Then, by matching the original cover text to the intercepted edition, one may find words that have been replaced. This alerts the interceptor not only of the intercepted message, but also of the information-hiding algorithm. Finally, the degree of occurrence of words on the same word list should be the same. Finding substitution classes is complex enough without having to include token frequencies. As frequency is considered in, it becomes even more complex.

In RNN-Stega, of Yang *et al.* [37] an artificial Steganographic text generation system based on RNN learns the statistical language distribution model from a vast volume of normal text until it generates messages that obey the statistical patterns. To achieve secret information masking, the conditional probability distribution of every word is coded using a binary tree throughout the sentence synthesis process. The approach, on the other hand, is useless for overall security since it only analyses the statistical features of a single phrase, neglecting the broader distribution of batch-generated texts. Din *et al.* [10] writes that the majority of Machine Translation systems used today is statistical MT particularly reliant on templates drawn from a corpus, as opposed to transfer systems based on linguistic principles. To construct a translation model in statistical MT, the algorithm is trained using a linguistic concurrent dataset. The translation model provides predictive information to the interpreter about possible word alignments. A word alignment is when the words in the source text and the destination text are in the same order. In Long *et al.* [1] since there are no modifications to the carrier, coverless information hiding can easily survive steganalysis and detection algorithms. Classic methods, on the other hand, have weaknesses such as weak hiding capabilities and a low hiding success quality. However, a text-coverless masking approach based on the word2vec method exceeds current techniques in terms of capacity. The approach is Word2vec, which is a natural language analysis tool. The word2vec programme learns word connections from a huge collection of text using a neural network model. After a phrase has been learnt, a model like this would distinguish synonyms and recommend further words. The text created by Li *et al.* [23] representation has the identical related meaning and subject as the knowledge graph since it encodes entities and relationships data from Knowledge Graphs. For the Steganography, graph structured data is used, which is then converted to text. Content based neural linguistic Steganography model used. The model generates Steganographic text for a given topic using information graph data. The model takes a secret message in the form of a bit-stream and converts it to Stego-text using a graph embedding network, which is based on transformer architecture. The statistical language model (RNN) used to generate Stego-text resulted in a significant number of regular text and a corpus of candidate words.

Fang *et al.* [12] Stego is AI-RNN architecture, which is a LSTM neural network. The model takes a hidden bit stream and a shared key and uses social media data mostly optimized for Twitter and Enron email data sets. An LSTM assembly consists of a cell, an input gate, an output gate, and a forget gate. The cell has three gates that control the flow of information in / out, and it recognises values for random timeframes. For classification, processing, and prediction, time series data is well-suited to LSTM networks.

Artificial Intelligence and Information Security: A Case Study Artificial Intelligence covers a broad range of topics, from Natural Language Processing, Neural Networks, Expert Systems, Fuzzy Logic, and Robotics. These fields have demonstrated promising results in speech and voice recognition. In the field of information technology, such as cryptography [8], [20] and Steganography, artificial intelligence is widely used.

The origins of artificial intelligence can be traced back to a short story written in 1940 about a robot called "Run around." Around that time, the Artificial Intelligence smelled

for the first time. Alan Turing, a British mathematician, wrote an essay titled ''Computing Machinery and Intelligence'' in 1950. Marvin Minsky and John McCarthy, a Stanford computer scientist, formally coined the term ''Artificial Intelligence'' in 1956. ELIZA, an AI computer program developed at MIT in 1964, was a natural language processing method that could converse with a person. Then, in 1997, expert systems emerged that could use a list of rules in a way that was similar to human intelligence. IBM's Deep Blue, a chess-playing computer, could, for example, analyse 200 million possible moves per second and determine the best next move. Gary Kasporov, the world champion, was defeated by this computer game. Expert systems, on the other hand, are unable to identify the images due to the system's lack of learning capabilities. In contrast to the process of neurons in the human brain, statistical methods were able to remember. Artificial Neural Networks technology was born as a result of this. Artificial Neural Networks became Deep Learning after a Google-developed algorithm, AlphaGo, defeated the world champion in the board game Go in 2015. Currently, Deep learning approaches are at the heart of visual recognition, speech recognition, Face book's natural language translations, and self-driving vehicles. It has now become a part of everyone's life in the form of Google search engines, which accept audio, text, and image inputs [27]. AI and deep learning systems have made this possible. Artificial intelligence (AI) systems are used in a variety of settings for decision-making. AI will assist in the automated machine-driven collection of goods, costs, website content, and promotional messages that are tailored to the needs of each particular user [18].

V. APPLICATIONS

Artificial Intelligence (AI) is advancing at a breakneck rate, presenting unparalleled opportunities to improve the success of various sectors and companies. The use of AI in the transportation domain aims to address issues such as rising travel demand, safety problems, emissions from fuel combustion, and pollution. The use of driverless vehicles can greatly reduce the number of collisions, as can the use of AI prediction and detection models to help forecast traffic flow. Google Maps is a powerful AI program that helps travellers by reducing traffic and predicting the shortest path, traffic frequency, and alternative routes [2].

In the past, making a correct dermatological diagnosis necessitated years of experience for thousands of patients. Artificial intelligence (AI) has made significant progress in current years, especially in the domain of image processing. As a result, computer scientists have developed algorithms that can identify skin lesions, especially melanoma, using these techniques [11], [19].

Artificial neural networks are a popular machine learning technique that is used in major disease fields such as cancer and cardiology, according to a new study of AI implementations in health care. Clinical evaluation, cancer forecasting, voice recognition, length of stay prediction, visual processing and understanding, and drug discovery are also examples of ANN applications in health care. Non-clinical uses include bettering health-care corporate management and forecasting primary metrics like cost and hospital use. ANN has been used in decision making frameworks to provide cost-effective alternatives to primary care providers and the health care system [29].

When it comes to evaluating images, image recognition and AI techniques will come in handy. Checking for COVID-19 positive cases currently depends heavily on Reverse Transcription-Polymerase Chain Reaction (RT-PCR) that is time intensive and has a false-negative error rate. As a result, it's critical to create novel methods for identifying patients more quickly and with greater precision. CT or X-Ray scans are one way to detect patients, but they require more readily available devices. By analyzing these images, it is possible to identify patients well before they experience symptoms such as fever or coughing [31], [40].

Artificial Intelligence in Education, according to numerous sources, is one of the most rapidly developing areas of educational technology. Although it has been around for a while, educators are also unsure how to use it for pedagogical purposes on a larger scale, and how it will have a favourable impact on teaching and learning in higher education [16]. Currently developing smart instructional design and interactive tools that use AI to deliver learning, testing, and feedback to students from elementary school to college level that provides them with the issues they're ready for, recognizes skill gaps, and redirects them to current concepts as required. If artificial intelligence advances, a computer may be able to recognise the feelings on a student's face as a sign that they are struggling with a concept and fine tune the lesson appropriately. Curricula that are tailored to each student's needs are not currently possible, but they will be in the future for AI systems [4].

Every day, network attacks get more nuanced and advanced. Aside from the so-called computer hackers and hacking novices, there are a slew of sophisticated criminals looking to cash in on corporate networks. In order to hack, intercept, or inflict harm more efficiently, oppressive states, large companies, and mafias are increasingly growing their expertise and abilities in cyber attacks. Orthodox methods to network security seem to be reaching their limits, and the need for a better strategy to threat identification is becoming apparent. Only a few methods are currently used in security implementations for decent performance, and they are limited to Machine Learning. Applications of Artificial Intelligence to Network Security have shown some promising findings using supervised machine learning. Until now, Unsupervised ML has seemed to be the subject of the majority of research in this field, as well as producing some impressive results and achieving levels of precision that were previously unattainable. Unsupervised ML, on the other hand, is also heavily reliant on human intelligence for meaning and information in order to take advantage of results. To create algorithms capable of delivering expert information rather than relying

on humans to recognize patterns found by Self-Organizing Maps or clustering techniques, the use of Bayesian Belief Networks to build expert systems is one of the ideas that seem to be gaining ground in this context. BNs, also known as Causal Probabilistic Networks, are a way of representing probabilistic interactions between various events. It is critical to keep focusing on data visualization to enhance security specialists' capacity to understand a wider spectrum of risks with less time, with less commitment and money. With the rise of Data Analytics, Big Data the ability to store and process massive volumes of data is becoming extremely relevant, and machine learning is playing a key part [13].

AI can be seen in a variety of ways in the area of cyber security of Industry 4.0. Identification and authentication, anomaly detection (such as apps, ransom ware, data stream anomalies, and so on), and threat evaluation are the core uses of AI in cyber security. AI algorithms are used for known attacks by the Intrusion Detection and Prevention System (IDPS). In this case, the algorithms are trained on the system's regular states. In this case, artificial intelligence may be used to classify threats and warnings in order to prevent false positive alarms triggered by irregular security staff actions [30].

A. AI CHALLENGES

The growing usage of the Internet and the everyday generation of massive amounts of data are posing new difficulties in the form of security vulnerabilities and breaches. A great deal of research is being done to discover answers to information security challenges, and one such area is the possible application of AI. The key objective is to develop effective security measures based on AI technologies. Previously, manual investigations were used to identify security flaws. However, due to the rise in security threats, the manual method is no longer capable of withstanding them. The greatest option is to use AI automated approaches, although adapting and implementing AI techniques is still in its infancy. There are many different types of information security risks that must be handled independently, and since they are so diverse, applying AI solutions to them would be more challenging, implying that practical AI deployment approaches would take time [3]. AI problems come in a variety of shapes and sizes. Individuals, small businesses are hesitant to use AI because of the significant costs involved. Human interaction is minimal due to their rapid and automated processes, resulting in job insecurity. Integration of AI and Big data approaches necessitates systems that can scale up as needed. To protect copyrights and ethics, considerable adjustments in current governmental, judicial, and ethical systems are required [21].

Information security refers to a system of procedures for protecting data against unwanted entry or modification. The terms ''information protection'' and ''cyber security'' are also used interchangeably. Information security encompasses the wider activity of protecting IT infrastructure from attack, and cyber security is a component of it. Computer technology has two sister practices: network security and application security (see Figure 4) [15].

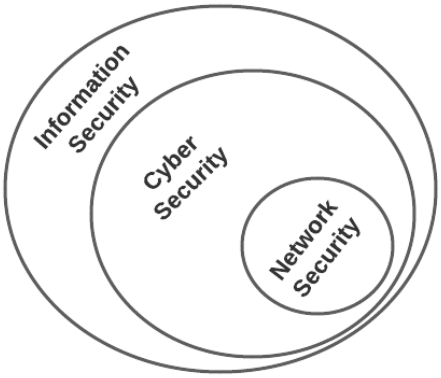

FIGURE 4. Linguistic steganography types.

B. AI AUTHENTICATION

AI Authentication is divided into two types: physical biometrics and behavioural biometrics. Physical biometry, as the name implies, is concerned with the identification of human bodily parts. Any individual may be enabled or disabled to access a lawful resource using AI facial identification, voice identification, and other applications. Behavioural biometry is based on distinct human patterns such as a person's typing speed, interaction, and motions, among others. Physical authentication is done solely at login time, whereas behavioural authentication is dependent on actions [25].

C. AI MALWARE ANALYSIS

The core notion of malware and any bio-entity activity is quite similar. As a result, developing new techniques to resistance malware attacks is simple. Because of the complexity of malware, artificial intelligence technologies are widely used, and researchers have had some success, proposes an AI-based malware analysis system that significantly depends on AI technologies to aid malware analysts [24].

D. NLP POTENTIALS

Rather than guarding against the danger, NLP techniques are required to recognize attack patterns. Prior to the attack, it serves as a fantastic awareness mechanism. Cyber attackers' typical attack behaviour employs the same lexical kinds, when developing APT (Advanced Persistent Threat) allowing security specialists to identify potentially dangerous websites. N-Gram analysis is one such approach used to detect attack patterns [24].

E. MULTIFACTOR AUTHENTICATION

Authentication using login credentials is commonly provided by Secured Systems, yet this is insufficient for today's transactions. To access a genuine resource, it requires multiple forms of authentication. MFA (Multi Factor Authentication), which is adaptive and AI-based, is utilized for that purpose. These systems may gather and evaluate background

information to provide a risk score, which may then be used to take appropriate action [17].

F. AI DETECTING PHISHING THREATS

Phishing attacks are the most common cybercrime because they take advantage of a user's vulnerability. Whale phishing targets wealthy individuals, Pharming targets a number of people at once, vishing target an entity appearing to be a respected organisation, and smishing targets suspicious URLs sent over SMS. Semantic Analysis is provided by NLP in order to identify such activities [9].

VI. NATURAL LANGUAGE PROCESSING IN STEGANOGRAPHY- MARKOV MODEL

NLP is a subfield of AI that deals with natural language interface between computers and humans. Stochastic text Steganography is regarded as an immune strategy against the multitude of steganalysis approaches, and may be effectively deployed utilizing a Markov Chain (MC) encoder/decoder paired with Huffman Coding (HC).

Statistical NLP is primarily used to model sentences. It's a probability distribution over individual words that can be articulated as follows:

$$\mathrm{p(S)} = \mathrm{p}\left(\mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3, \ldots, \mathrm{v}_\mathrm{n}\right) \tag{1}$$

$$\mathrm{p(S)} = \mathrm{p}\left(\mathrm{v}_1\right)\mathrm{p}\left(\mathrm{v}_2 \mid \mathrm{v}_1\right)\ldots\mathrm{p}\left(\mathrm{v}_\mathrm{n} \mid \mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_{\mathrm{n}-1}\right) \tag{2}$$

where S refers to a sentence of word length n and $v_1, v_2, v_3, \ldots, v_n$ denotes individual terms. The likelihood of the sentence is p(S). When the first n -1 word is given, it is made up of the product of n conditional probabilities, each of which calculates the probability distribution of the nth word. As a result, in order to produce high-quality texts dynamically, we need a reasonable approximation of the training set statistical language model.

Markov chains, named after Andrey Markov, are mathematical constructs that move from one "state" to the next. Figure 5 is a straightforward two-state Markov chain. There are four potential transformations in state space with two states (A and B). We may either move to 'B' or remain at 'A' while we're at 'A.' We may either move to 'A' or remain at 'A' if we're at 'A.' The probability of transitioning from one state to the next is 0.5 in this two-state diagram.

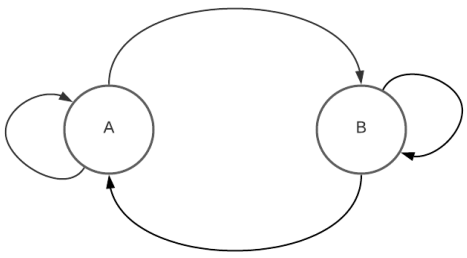

FIGURE 5. Two state Markov chain.

A Markov chain is a stochastic model in probability theory that describes a series of potential events in which the probability of each occurrence is solely dependent on the state achieved in the previous event. The Markov chain model is well suited to time series signal modelling. The Markov chain model can be expressed as follows: where x is the value space, q is the stochastic (random) variable set, and s is the sample space with a given probability distribution.

$$P\left(q_t = x^t\right) = f\left(P\left(q_{t-1} = x^{t-1}\right), P\left(q_{t-2} = x^{t-2}\right), \ldots, P\left(q_1 = x^1\right)\right) \tag{3}$$

such that,

$$\sum_{i=1}^{m} P\left(q_t = x_i\right) = 1, \quad \forall x_i \in N \tag{4}$$

The probability transfer function is denoted by f. The probability of the entire sequence can then be represented as:

$$\mathrm{P(Q)} = \mathrm{p}\left(\mathrm{q}_1, \mathrm{q}_2, \mathrm{q}_3, \ldots, \mathrm{q}_\mathrm{n}\right) \tag{5}$$

$$\mathrm{P(Q)} = \mathrm{p}\left(\mathrm{q}_1 = \mathrm{x}^1\right)\mathrm{p}\left(\mathrm{q}_2 = \mathrm{x}^2\right)\ldots\mathrm{p}\left(\mathrm{q}_\mathrm{n} = \mathrm{x}^\mathrm{n}\right) \tag{6}$$

$$\mathrm{P(Q)} = \mathrm{p}\left(\mathrm{q}_1\right)\mathrm{p}\left(\mathrm{q}_2 \mid \mathrm{q}_1\right)\ldots\mathrm{p}\left(\mathrm{q}_\mathrm{n} \mid \mathrm{q}_{\mathrm{n}-2}, \ldots, \mathrm{q}_1\right) \tag{7}$$

Since x^i represents the i^{th} word in the text in the above equations (1) and (3), it can rightly model the statistical language model of the text. Due to the above similarities, the Markov chain model is well suited to modelling text and is extensively used in natural language processing, especially in the field of automatic text generation. The generation of automatic text using a Markov model necessitates a large sample data set as well as the creation of a strong statistical language model using a dictionary of terms based on the training requirements. The following is a large dictionary; D contains all the words appeared in the training set.

$$\mathrm{D} = \left(\text{ word }_{\mathrm{D}_1}, \text{word }_{\mathrm{D}_2}, \text{word }_{\mathrm{D}_3}, \ldots, \text{ word }_{\mathrm{D}_\mathrm{N}}\right) \tag{8}$$

where, $word_{\mathrm{D}_1}$, represents, the $i^t h$ word in the dictionary D and N represents the total number of words.

The Figure 6 depicts the method of embedding hidden text using the Markov model and Huffman coding. This necessitates the provision of hidden text in bit stream format for embedding in the top layer. The Markov chain model is used to automatically produce texts; each time a word is created, the model determines the probability distribution $\mathrm{P}(word_{\mathrm{D}_1}|word_{\mathrm{S}_1}, word_{\mathrm{S}_1}, \ldots, word_{\mathrm{S}_t})$ of the following word, based on all of the words generated in the previous steps. Interpret all of the words in the dictionary D based on their conditional probability distribution, later choose the correct word based on the hidden bit stream to attain the objective of hiding the content. At each time point where the number of sentences in the learning sample set is large enough, there may be more than one possible solution. After descending the estimation likelihood of all the words in the dictionary D, we can select the top m sorted terms to create the Candidate Pool (CP). If the size of the candidate pool

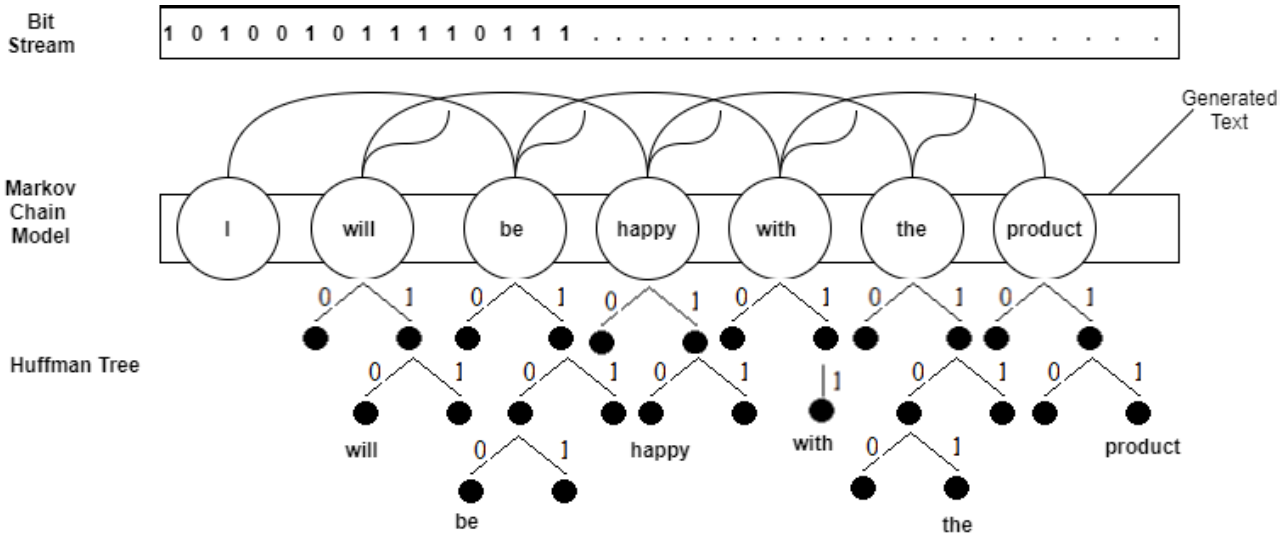

FIGURE 6. Embedding confidential text using Markov model and Huffman coding.

is appropriately chosen, the reliability of the produced text is always good. The Markov chain model automatically calculates the conditional probability distribution of the next letter when the first keyword "I" is entered, as seen in the diagram. After descending the likelihood of every word in the dictionary D, all the other necessary words are stored in the Candidate Pool. The end result was a fine, substantive Stego text. If the Candidate Pool is complete, the process of encoding the words begins; in this case, Huffman coding is used.

For all of the words in the Candidate Pool, Huffman tree coding is used to produce an optimal prefix code, which can be interpreted as a variable length code table. Every other word in the Candidate Pool is defined by a leaf node of the tree in the text generation process, the edges bind each non-leaf node, and its two child nodes are then encoded with 0 and 1, with 0 on the left and 1 on the right, as seen in Figure 6. The mechanism of information embedding is to choose the corresponding leaf node as the output of the current stage based on the binary code stream that needs to be embedded. A keyword list was used to eliminate the requirement that two identical sequences of bits yield two similar text sentences [22]. The probable terms were ordered descending by frequency. During the development of Stego text, the words were chosen at random. The receiver on the other hand receives the Stego text and reverses the process to obtain the hidden code. Both parties must commit to use the same methods in order to extract message properly. Flowcharts of execution methods express in Figure 7.

VII. ALGORITHM FOR MESSAGE EXTRACTION

Input: Stego Text, Buffer space for Dictionary, D and Candidate Pool, CP Output: Extracted Message

1) At the receiver end, the extraction procedure begins with gathering Stego text from the sender, which contains the message.
2) The probability distribution of each word in the Stego Text is calculated using a Markov chain and kept in a dictionary, D.
3) On the dictionary, the probability distribution of words is sorted in descending order and placed in the Candidate pool (CP).
4) For each word, create a Huffman tree and encode bit 0 to the left node and bit 1 to the right node.
5) Determine each word's root and leaf node.
6) Decode bits concealed in all the words
7) Finally, the message bit stream is produced as the resulting output, and the message is formed.

VIII. COMPARISON OF TWO RNN STEGANOGRAPHY METHODS

We compared two RNN, RNN-Stega [10] and RNN-generated Lyrics, in this study. The two methods were contrasted in terms of the language used, the form of training data set, the amount of training data, the methods, the coding system, embedding rate and capability, security, and other factors. The Table 1 below illustrates the specifics of the above-mentioned points in relation to two different strategies [28].

IX. EMBEDDING RATE AND CAPACITY USING RNN-STEGA

RNN-Stega is a form of linguistic Steganography that uses auto-generated Stego text as a secret message carrier. The recurrent neural network, or RNN, is used in this system. The key goal of this approach is to create natural-looking carrier texts for use in Steganography. Fixed-length coding (FLC) and variable-length coding (VLC) are the two methods proposed. Words are encoded using Huffman coding on the conditional probability distribution. As an RNN variant, the LSTM model is used. English social media info,

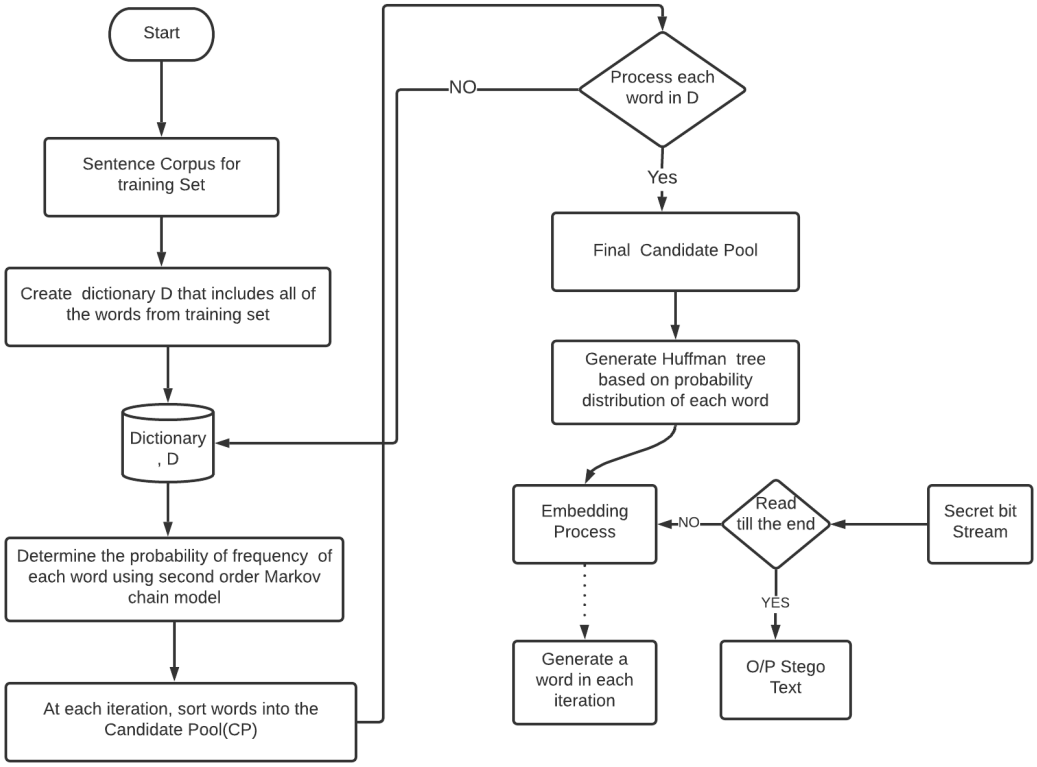

FIGURE 7. Flowcharts of execution methods.

TABLE 1. Comparison of two recurrent neural network steganographic methods.

Parameters	RNN-Stega	RNN-Generated lyrics
Language tested	English	Chinese
Training Data Set	Twitter messages, News, Movie Reviews	Chinese pop Music Lyrics
Confidential Message Used as Binary Stream	Used as Binary Stream	Used as Binary Stream
Candidate Pool Building	Using words	Using characters/words
Training Data set Volume	The New York Times, Breitbart, and CNN were among the 15 American outlets that contributed 143,000 stories. The Twitter API was used to extract 1,600,000 tweets.	15,000 Chinese pop lyrics from music websites, including 100 Chinese male singers, 100 Chinese female singers, 100 Chinese bands, and 50 popular songs of each singer.
Methods Tested	FLC, VLC, and Tina's Model	Char-RNN, Word-RNN, Poem, and Ci
RNN Variant	LSTM	LSTM
Coding Method	Based on Huffman Tree	Based on Huffman tree
Stego-Text Generation	Automatic	Automatic
Improvement	The FLC framework has a clear edge in terms of hiding data efficiency.	Used Char-RNN Model to build Word-RNN with improvement in the accuracy of Stego-text Generation

such as Twitter posts, news, and movie reviews, were used as the data collection [5].

Tina's model, which is a similar type of method, is contrasted to the proposed models, FLC and VLC. The rate and ability of information embedding are taken into account in this study. The three models previously described are compared in terms of embedding capability. The proposed model VLC and Tina's model are almost identical in terms of performance. Since each iteration of the model uses Huffman tree construction, which takes a long time. The FLC model, on the other hand, outperforms the other two. The conditional probability distribution with respect to a mapping from binary bit stream to word space is used in information hiding based on words. In each iteration, candidate pool space (CPS) is generated to produce a certain feasible word. The embedding potential increases as the amount of the CPS grows.

The FLC system employs a plain binary tree for coding, while the VLC and Tina's models both employ the Huffman tree. Because each repetition necessitates the generation of a Huffman tree, which takes a long time, producing a 50 word

text takes about 46 seconds. On the other hand, the FLC approach takes about 3 seconds to generate the same 50 word text, indicating good information hiding performance. The Figure 8 & Table 2 depicts the same thing.

TABLE 2. Comparison of embedding rate of three methods (FLC, VLC, & TINA) using RNN text steganography.

Candidate Pool Space (No. of words)	FLC	VLC	TINA
1	3	3	3
2	4	4.5	7
4	5	8.5	9
7	5	14	17
16	5	25	29
33	5	48	47.5

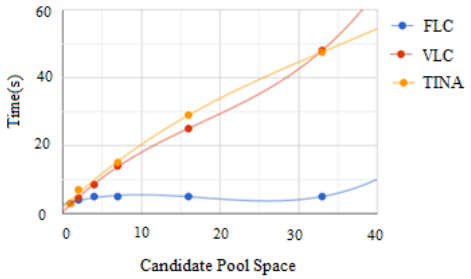

FIGURE 8. Graph showing comparison of embedding rate of three methods (FLC, VLC, and Tina) using RNN text steganography.

X. EMBEDDING CAPACITY AND SECURITY OF RNN-GENERATED LYRICS

Chinese pop lyrics are used as the sample data set for the RNN-generated lyrics method. The proposed system Word-RNN is a refinement of the current Char-RNN form. The relative Stego text in the form of Chinese lyrics is created automatically based on the binary stream of secret information length. For producing lyrics, an LSTM model with a Huffman tree is used to conceal the hidden information.

As compared to other methods such as Poem-based Steganography, Ci-based Steganography, and Char-RNN, this approach Word-RNN demonstrates an increase in embedding capability. All of the Steganography approaches are tailored to the Chinese language. The embedding result relation for each of the four processes is shown in tabular form of Table 3. The size of the candidate pool, which is the total number of candidate terms that can be accommodated, is a significant factor in embedding ability. More confidential knowledge may be embedded in a larger candidate pool.

It is clear from the table that the Word-RNN and Char-RNN have higher embedding power with the increase in Candidate pool size. Since the lyrics are longer, they have a higher embedding power. Word-RNN and Char-RNN have greater embedding capability than the other two methods, Poem and Ci, as seen in the graph (Figure 9) above. These methods' protection is based on arousing scepticism during the text carrier

TABLE 3. The data extracted from [42] for four methods with respect to the embedding capacity with the varying size of candidate pool.

Size. of Candidate pool	Embedding capacity			
	Poem Method	Ci Method	Char-RNN	Word-RNN
2	0	0	10	8
4	2	5	15	10
8	4	10	40	18
16	6	10	45	20
32	8	12	50	40
64	8	19	70	36

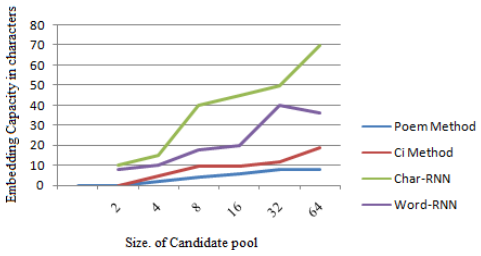

FIGURE 9. Comparison of embedding capacity of four methods (Poem, Ci, Char-RNN, & Word-RNN).

distribution period. Poems and modern popular lyrics play an important role in Chinese language. Traditional methods of generating poetry may raise doubt, but modern artificial intelligence methods such as RNN Steganography are very effective at generating lyrics that are more natural. As a result, the proposed approaches are safer than the previous ones.

XI. GAPS IDENTIFIED IN THE CURRENT STUDY

Conventional Steganographic methods are inefficient in every way, yet a touch of artificial intelligence produces excellent results. AI approaches are now being applied for both positive and negative steganography. According to the present trend, more study is being conducted on steganalysis rather than data concealing, which is a source of concern. Application of Deep learning in steganography is promising, but it takes more processing power and runs slower. Data concealing using a Recurrent Neural Network is more efficient in terms of payload capacity than traditional approaches; nevertheless, training an RNN is complex, and processing long words/sequences is problematic. Data concealing qualities are frequently linked, with improving one lowering the value of the other. As a result, striking the optimal balance between payload capacity, undetectability, and robustness is extremely difficult. AI-based solutions have gained momentum over the last ten years, but they are still in their infancy. Traditional approaches are less effective than AI-based solutions. Machine learning, deep learning, and other NLP technologies are needed to enhance traditional approaches. Integrating cryptography with steganography is critical for improving security. Standards for Steganography, like Cryptography, must be established. There is a need for further Steganographic study in languages other than English.

The study emphasises the embedding rate and Embedding capacity of RNN-based Steganographic techniques evaluated

in the last portion of the comparison. As the quantity of Candidate Pool Space grows, so does the embedding potential. Due to the production of a Huffman tree for each repeat, embedding with a Huffman tree takes longer than embedding with a plain binary tree. The word-RNN and character-based RNN techniques have a higher embedding capacity than the poem-based and Ci-based Steganography methods. Because of the popularity of poems and lyrics as Stego material, the security of the traditional techniques stated above is often questioned.

XII. CONCLUSION

This article evaluates the advantages of artificial intelligence techniques in linguistic Steganography in terms of adding automatic functionality to existing methods to increase hiding ability and rate, reduce suspicion of auto-generated Stego, and enhance security. For most conventional linguistic Steganographic approaches, poor embedding capabilities may be a problem. Linguistic Steganography based on text generation, on the other hand, effectively solves this issue (this fact is illustrated in the comparison). The Markov chain based Steganography model can be used to counter third-party attacks viewed on statistical properties of natural language, as well as to produce the best possible sentences. However, owing to the Markov model's constraints, the text given by the Markov model lacks accuracy, making it difficult to recognise. As NLP technology has progressed, a growing number of Steganographic text generation models based on neural network models have emerged in recent years. Recent studies have shown that, even if the resulting text is precise sufficiently, current models cannot monitor the semantics (meaning) of generated Steganographic texts, posing a security risk. Analyzing the technology for automatic Steganographic text formation with controllable semantics is a problem that needs to be solved.

XIII. CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper References.

REFERENCES

[1] *Text Coverless Information Hiding Based on Word2vec*. Springer Professional.
[2] R. Abduljabbar, H. Dia, S. Liyanage, and S. A. Bagloee, "Applications of artificial intelligence in transport: An overview," *Sustainability*, vol. 11, no. 1, p. 189, Jan. 2019.
[3] R. R. Althar and D. Samanta, "The realist approach for evaluation of computational intelligence in software engineering," *Innov. Syst. Softw. Eng.*, vol. 17, no. 1, pp. 17–27, Mar. 2021.
[4] A. K. Biswal, D. Singh, B. K. Pattanayak, D. Samanta, and M.-H. Yang, "IoT-based smart alert system for drowsy driver detection," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–13, Mar. 2021.
[5] J. Biswas, P. Kayal, and D. Samanta, "Reducing approximation error with rapid convergence rate for non-negative matrix factorization (NMF)," *Math. Statist.*, vol. 9, no. 3, pp. 285–289, May 2021.
[6] C. Chang and S. Clark, "Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method," *Comput. Linguistics*, vol. 40, no. 2, pp. 403–448, Jun. 2014.
[7] R. Gurunath and D. Samanta, "Studies on encrypted secret data storage techniques analogous to steganography," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 2, pp. 3705–3711, Jan. 2020.
[8] S. K. A. Khadri, D. Samanta, and M. Paul, "Message encryption using text inversion plus n count: In cryptology," *Int. J. Inf. Sci. Intell. Syst.*, vol. 3, no. 2, pp. 71–74, 2014.
[9] V. Dhanush, A. R. Mahendra, M. V. Kumudavalli, and D. Samanta, "Application of deep learning technique for automatic data exchange with air-gapped systems and its security concerns," in *Proc. Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Jul. 2017, pp. 324–328.
[10] R. Din, S. Utama, and A. Mustapha, "Evaluation review on effectiveness and security performances of text steganography technique," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 11, no. 2, pp. 747–754, Aug. 2018.
[11] X. Du-Harpur, F. M. Watt, N. M. Luscombe, and M. D. Lynch, "What is AI? Applications of artificial intelligence to dermatology," *Brit. J. Dermatol.*, vol. 183, no. 3, pp. 423–430, Sep. 2020.
[12] T. Fang, M. Jaggi, and K. Argyraki, "Generating steganographic text with LSTMs," May 2017, *arXiv:1705.10742*. [Online]. Available: http://arxiv.org/abs/1705.10742
[13] V. Gomathy, N. Padhy, D. Samanta, M. Sivaram, V. Jain, and I. S. Amiri, "Malicious node detection using heterogeneous cluster based secure routing protocol (HCBS) in wireless adhoc sensor networks," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 11, pp. 4995–5001, Nov. 2020.
[14] M. Grosvald and C. O. Orgun, "Free from the cover text: A human-generated natural language approach to text-based steganography," *J. Inf. Hiding Multimedia Signal Process.*, vol. 2, no. 2, pp. 1–9, 2011.
[15] A. Guha and D. Samanta, "Hybrid approach to document anomaly detection: An application to facilitate RPA in title insurance," *Int. J. Autom. Comput.*, vol. 18, no. 1, pp. 55–72, Feb. 2021.
[16] A. Guha, D. Samanta, A. Banerjee, and D. Agarwal, "A deep learning model for information loss prevention from multi-page digital documents," *IEEE Access*, vol. 9, pp. 80451–80465, 2021.
[17] R. Gurunath, M. Agarwal, A. Nandi, and D. Samanta, "An overview: Security issue in IoT network," in *Proc. 2nd Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Aug. 2018, pp. 104–107.
[18] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," *California Manage. Rev.*, vol. 61, no. 4, pp. 5–14, Aug. 2019.
[19] K. N. Pradhan, M. Siddappa, S. Kavitha, and D. Samanta, "Analysis & improvement of wireless network security based on biometrics," Tech. Rep., Mar. 2019.
[20] S. K. A. Khadri, D. Samanta, and M. Paul, "Message encryption using Pascal triangle multiplication: In cryptology," *Asian J. Math. Comput. Res.*, vol. 13, pp. 262–270, Sep. 2016.
[21] A. Khamparia, P. K. Singh, P. Rani, D. Samanta, A. Khanna, and B. Bhushan, "An Internet of Health Things-driven deep learning framework for detection and classification of skin cancer using transfer learning," *Trans. Emerg. Telecommun. Technol.*, May 2020.
[22] R. Kumar, R. Kumar, D. Samanta, M. Paul, and V. Kumar, "A combining approach using DFT and FIR filter to enhance impulse response," in *Proc. Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Jul. 2017, pp. 134–137.
[23] Y. Li, J. Zhang, Z. Yang, and R. Zhang, "Topic-aware neural linguistic steganography based on knowledge graphs," *ACM/IMS Trans. Data Sci.*, vol. 2, no. 2, pp. 10:1–10:13, Apr. 2021.
[24] M. Maheswari, S. Geetha, S. S. Kumar, M. Karuppiah, D. Samanta, and Y. Park, "PEVRM: Probabilistic evolution based version recommendation model for mobile applications," *IEEE Access*, vol. 9, pp. 20819–20827, 2021.
[25] M. S. Mekala, R. Patan, S. K. H. Islam, D. Samanta, G. A. Mallah, and S. A. Chaudhry, "DAWM: Cost-aware asset claim analysis approach on big data analytic computation model for cloud data centre," Tech. Rep., May 2021.
[26] L. Y. Por and B. Delina, "Information hiding: A new approach in text," Tech. Rep., 2008.
[27] S. Pramanik, D. Samanta, S. Dutta, R. Ghosh, M. Ghonge, and D. Pandey, "Steganography using improved LSB approach and asymmetric cryptography," in *Proc. IEEE Int. Conf. Advent Trends Multidisciplinary Res. Innov. (ICATMRI)*, Dec. 2020, pp. 1–5.
[28] D. Samanta, M. G. Galety, M. Shivamurthaiah, and S. Kariyappala, "A hybridization approach based semantic approach to the software engineering," *TEST Eng. Manage.*, vol. 83, pp. 5441–5447, Mar. 2020.
[29] N. Shahid, T. Rappon, and W. Berta, "Applications of artificial neural networks in health care organizational decision-making: A scoping review," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0212356.

[30] P. Sivakumar, R. Nagaraju, D. Samanta, M. Sivaram, M. N. Hindia, and I. S. Amiri, "A novel free space communication system using nonlinear InGaAsP microsystem resonators for enabling power-control toward smart cities," *Wireless Netw.*, vol. 26, no. 4, pp. 2317–2328, May 2020.
[31] M.-H. N. Tayarani, "Applications of artificial intelligence in battling against COVID-19: A literature review," *Chaos, Solitons Fractals*, vol. 142, Jan. 2021, Art. no. 110338.
[32] M. Y. Valandar, P. Ayubi, and M. J. Barani, "A new transform domain steganography based on modified logistic chaotic map for color images," *J. Inf. Secur. Appl.*, vol. 34, pp. 142–151, Jun. 2017.
[33] M. Y. Valandar, M. J. Barani, P. Ayubi, and M. Aghazadeh, "An integer wavelet transform image steganography method based on 3D sine chaotic map," *Multimedia Tools Appl.*, vol. 78, no. 8, pp. 9971–9989, 2019.
[34] N. Wu, P. Shang, J. Fan, Z. Yang, W. Ma, and Z. Liu, "Research on coverless text steganography based on single bit rules," *J. Phys., Conf. Ser.*, vol. 1237, Jun. 2019, Art. no. 022077.
[35] L. Xiang, S. Yang, Y. Liu, Q. Li, and C. Zhu, "Novel linguistic steganography based on character-level text generation," *Mathematics*, vol. 8, no. 9, p. 1558, Sep. 2020.
[36] S. R. Yaghobi and H. Sajedi, "Text steganography in webometrics," *Int. J. Inf. Technol.*, vol. 13, no. 2, pp. 621–635, Apr. 2021.
[37] Z. Yang, X. Guo, Z. Chen, Y. Huang, and Y. Zhang, "RNN-Stega: Linguistic steganography based on recurrent neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1280–1295, May 2019.
[38] Z. Yang, S. Jin, Y. Huang, Y. Zhang, and H. Li, "Automatically generate steganographic text based on Markov model and Huffman coding," Nov. 2018, *arXiv:1811.04720*. [Online]. Available: http://arxiv.org/abs/1811.04720
[39] Z. Yang, L. Xiang, S. Zhang, X. Sun, and Y. Huang, "Linguistic generative steganography with enhanced cognitive-imperceptibility," *IEEE Signal Process. Lett.*, vol. 28, pp. 409–413, 2021.
[40] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education—Where are the educators?" *Int. J. Educ. Technol. Higher Educ.*, vol. 16, no. 1, p. 39, Oct. 2019.

R. GURUNATH is currently an Assistant Professor with Dayananda Sagar College of Engineering, Bengaluru, India, and a Research Scholar with the Department of Computer Science, CHRIST (Deemed to be University), India. His research interest includes text steganography.

AHMED H. ALAHMADI received the Ph.D. degree in computer science and engineering from La Trobe University. His Ph.D. research was in e-health business requirements engineering. Since then, he has published various peer-reviewed research articles. He worked as the Dean of the College of Computer Science and IT, Albaha University. He also has a demonstrated history of working in the higher education industry. He is currently an Assistant Professor with the Department of Computer Science and Information, Taibah University, Saudi Arabia. He is also the Dean of Khaybar Community College, Taibah University. In addition to research, he is also skilled in accreditation and college recruiting. He has made significant contributions in various research areas, including e-health, software engineering, business process modeling, requirements engineering, and process mining.

DEBABRATA SAMANTA (Member, IEEE) received the bachelor's degree (Hons.) in physics from the University of Calcutta, Kolkata, India, the M.C.A. degree from the Academy of Technology under WBUT, West Bengal, and the Ph.D. degree in computer science and engineering from the National Institute of Technology Durgapur, Durgapur, India, in the area of SAR image processing. He is currently working as an Assistant Professor with the Department of Computer Science, CHRIST (Deemed to be University), Bengaluru, India. He is keenly interested in interdisciplinary research and development and has experience spanning fields of SAR image analysis, video surveillance, heuristic algorithm for image classification, deep learning framework for detection and classification, blockchain, statistical modeling, wireless *ad-hoc* networks, natural language processing, and V2I communication. He has successfully completed six consultancy projects. He is the owner of 18 patents (two design Indian patent and two Australian patent granted, and 14 Indian patent published) and two copyrights. He has authored or coauthored over 151 research papers in international journal (SCI/SCIE/ESCI/Scopus) and conferences, including IEEE, Springer, and Elsevier conference proceeding. He has also coauthored ten books and coedited five books, available for sale on Amazon and Flipkart. He has authored or coauthored 19 book chapters. He is a Professional IEEE Member, an Associate Life Member of Computer Society of India (CSI), and a Life Member of the Indian Society for Technical Education (ISTE). He has received the Scholastic Award at 2nd International conference on Computer Science and IT Application, CSIT-2011, Delhi, India. He has received funding under the International Travel Support Scheme, in 2019, for attending conference in Thailand. He has received Travel Grant for speaker in conference and seminar, for two years, in July 2019. He has presented various papers at international conferences and received best paper awards. He is a convener, a keynote speaker, the session chair, the co-chair, the publicity chair, the publication chair, and the advisory board and technical program committee member in many prestigious international and national conferences. He was an invited speaker at several institutions. He serves as an Acquisition Editor for Springer, Wiley, CRC, Scrivener Publishing LLC, Beverly, USA, and Elsevier.

MOHAMMAD ZUBAIR KHAN received the M.Tech. degree in computer science and engineering from Uttar Pradesh Technical University, Lucknow, India, in 2006, and the Ph.D. degree in computer science and information technology from the Faculty of Engineering, Mahatma Jyotiba Phule Rohilkhand University, Bareilly, India. He was the Head and an Associate Professor with the Department of Computer Science and Engineering, Invertis University Bareilly. He has more than 15 years of teaching and research experience. He is currently an Associate Professor with the Department of Computer Science, Taibah University. He has published more than 60 journals and conference papers. His current research interests include data mining, big data, parallel and distributed computing, theory of computations, and computer networks. He has been a member of the Computer Society of India, since 2004.

ABDULRAHMAN ALAHMADI received the Ph.D. degree in computer science and engineering from Southern Illinois University at Carbondale, Carbondale, in 2019. His Ph.D. research was in cloud computing data centers scheduling for energy consumption reduction and resource utilization improvement. During his studies, he was working with the Cloud Computing and Big Data Research Laboratory, for five years. He is currently an Assistant Professor with the Department of Computer Science and Information, Taibah University, Saudi Arabia. Since then, he has published various peer-reviewed research papers in edge and fog cloud computing. His research interests include machine learning resource management in cloud computing, task scheduling in fog computing, and the IoT-supported edge offloading techniques.