Journal of Modern Computing and Engineering Research
JMCER
Volume 2024, p. 1-8
https://jmcer.org

# Proposed Method for Text Steganography

Alaa Abdullah Idres[1] and Yaseen Hikmat Ismael[2]

[1, 2] *Department of Computer Science, College of Computer Science and Mathematics, University of Mosul*

[1]*alaa.22csp37@student.uomosul.edu.iq*, [2]*yaseen-hikmat@uomosul.edu.iq*

*Received: November 30, 2023* *Revised: February 8, 2024* *Accepted: February 10, 2024*

**Abstract**

Many people can now communicate with each other easily and at a high speed. However, using the Internet for communication is accompanied by the problem of providing protection. Researchers initially used various encryption systems to provide data protection, but seeing the data in its encrypted form raises suspicion in the attacker or intruder that significant and sensitive data has been encrypted, which uses different methods to break the code and try to find out its content. Information-hiding techniques have emerged as a tool embedded in secret data within a transmission medium so the attacker will not notice the presence of that data. Steganography technology uses different media types to cover includes the secret message. The media can be an images, audio, video, text, or other media. The process of hiding in text involves a set of challenges or difficulties, the most important of which is the lack of spaces that can be exploited in the process of hiding, in addition to the significant influence of the text on the hidden data. This research proposes a method for hiding inside the text. The technique consists of two levels. The first is searching for the letters of the secret message in the cover text and encoding those locations. In contrast, the second level includes converting the symbols of the letter locations into a binary form and hiding them in the cover text in a new way. The proposed method achieved a high level of security for hiding data within the text, making it difficult for an attacker to discover it.

**Keywords:** *Text-Steganography, Information Hiding, Text Confidentiality.*

## 1 Introduction

Due to the suspicion raised by various encryption methods used in data protection, particularly when sensitive and important data is involved, information steganography has emerged as a fundamental and essential alternative for preserving data security. The process of data steganography involves using a specific medium as a cover to embed secret information. There are various types of media (such as images, audio, video, and text). Because of the limited space available for embedding secret data when using text as a cover, as well as the significant sensitivity of text to changes, text-based data steganography has garnered significant attention among researchers. Many studies and techniques have emerged in this field. Text-based data steganography technology can be divided into three main types: format-based, random and statistical generation, and linguistic methods (Sun, B. et al., 2023).

## 2 Text Steganography Techniques

Text steganography techniques are used to explain the techniques and methods used in hiding within the text, as well as identify the most important advantages of these techniques (Akbar, F.C. and et al., 2020; Hamdan, A.M. and Hamarsheh, A. 2016).

### *2.1 Format-based method*

The present method includes using methods based on changing the text format to hide data, such as changing font sizes or adding white distances between words. One of the disadvantages of this method is that if the STEGO file is opened using the text processor, the spelling errors and white distances (the spaces added in the concealment process) will be recognized. In addition, if a comparison is made between the original text (before concealment) and the text after hiding, the places of change will be recognized, thus increasing the possibility of discovering the hidden text [2, 3].

*Corresponding author: alaa.22csp37@student.uomosul.edu.iq*
P-ISSN: 2960-141X E-ISSN: 2958-0021

JMCER
JOURNAL OF MODERN COMPUTING AND ENGINEERING RESEARCH

2.1.1 Line Shift

Line shift means that the distances between the lines of the text are fixed. In this method, for the purpose of hiding the text, the text is initially converted into the values of the ASCII coding and then converted into the binary formula. For the purpose of hiding the bit (O), the line is removed to the top with a fixed value (the amount of a fixed displacement), while the concealment is the value of (1). The line is removed down and with the same amount of displacement upwards. One of the disadvantages of this method is that if the text (edited) is rewritten using a different text processor or using the letters of the alphabet (OCR), hidden information will be destroyed (Akbar, F.C. et al., 2020). The line shift can be clarified in Fig. 1.

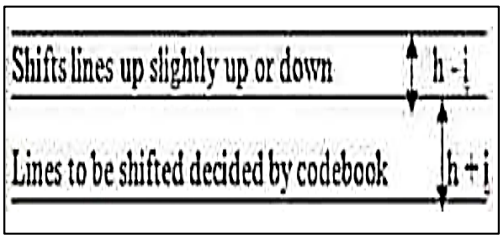


Figure 1: Line shift text steganography.

2.1.2 Word Shift

The technique for word shifting relies on altering the ordinary text's word spacing. Throughout the text, there is constant word spacing. The word is moved to the right by a predetermined amount to conceal the value (0) and to the left by a predetermined amount to conceal the value (1), as agreed upon by the two communication parties. While this method is thought to be simple for masking, one drawback is that it is simple to identify the masking process by the use of an optical character recognition application, a different text editor, or by comparing the original text with the masking text (Akbar, F.C. and et al., 2020). The word shift method can be clarified in Fig. 2.

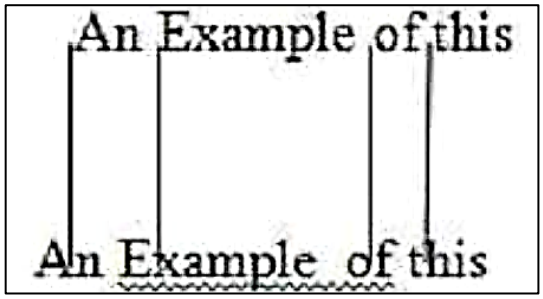


Figure 2. Word shift text steganography.

2.1.3 Feature Coding

The method of feature coding is used to perform the concealment inside the text by changing some of the features of the text letters. For example, to hide the value (0), the last letter of the word (which extends its end) is extended, and to hide the value (1), the end of the last letter of the word is shortened. One of the advantages of this method is that a large hidden space can be provided, but its faults, as is the case in all the methods of the format-based method, can be discovered in the concealment process in the same previous methods (Mandal, K.K. and Singh P.K., 2019).

## *2.2 Random and Statistical Generation*

It is used when a random text is generated by relying on statistical characteristics to include the secret text (the concealment process). One of the methods used in this method is to hide the data in a random display sequence for texts or use statistical features of the word length and frequency of letters to produce words that have statistical properties similar to actual words in a specific language. One of the most important features of random statistical generation is the lack of original text (before concealment), which the attacker compares to discover the concealment (Khan, Y. et al., 2021).

2.2.1 Word Mapping

This technique converts the secret message to be hidden using the cross-over process, after which the resulting encoded text appears on the cover. The concealment process depends on the length of the floor in the cover file, whether it is my husband or an individual. The hiding site map is sent in a separate file from the concealment file to the receiving side (Agarwal, M., 2013).

2.2.2 MS Word Document

This method uses the different formatting features of Microsoft Word, where the text appears after the concealment process as if it is a result of writing a person with a few skills using the Microsoft Word program. The use of various formats to carry out the concealment process depends on the agreement between the two sides of the communication (Agarwal, M., 2013).

### *2.3 Linguistic Methods*

The linguistic method depends on the linguistic characteristics of the text that have been created and changed by the concealment. For example, CFG Grammar is a group of rules that produce a possible tree or analysis and thus rely on it to discover the hidden text. Where the left branch is determined by (0) and the right branch by (1). Greibach Normal Form (GNF) can also be used to generate a brown tree to achieve concealment. One of the disadvantages of this method is that the rules of small grammar lead to repeating a text clip several times. In addition, although the text is not tainted with grammatical flaws, the semantic structure of the text is a series of sentences that do not have a link to each other (Majeed, M.A. et al., 2021).

#### 2.3.1 Syntactic Method

The syntactic method of the text depends on the use of the point (.), the comma (,), the decisive comma (;), and so on to hide the values of bits (1,0). The problem with this method is that it requires determining the right places to include numbering marks. The wrong use and punctuation marks in incorrect places lead to the ease of detecting the concept (Taleby A. et al., 2019).

#### 2.3.2 Semantic Method

The semantic method is used as the linguistic synonym of certain words and thus the concealment process in the locations of these synonyms. One of the disadvantages of this method is the limited concealment process, as well as the fact that this concealment process may completely change the meaning of the text (Taleby A. et al., 2019).

## 3 Related Work

Chaw A. has presented a new way to hide inside the text. He depended on the use of the semantic method, in replacement words of the cover of the cover with the words synonymous with it. After converting the text to be hidden to the ASCII system and then to the binary system. In instances where synonyms are used, the value (Zero, Ones) is concealed. The researcher applied the suggested concealment technique using the system as a model for exchanging banking information between banks and clients (Chaw, A.A., 2019).

Al-Nofaie S. et al. recommended a way to conceal the texts written in Arabic or similar languages such as Persian and Urdu. For the purpose of performing a safe concealment, the researchers used the fake distance (PS-BetWord) and (Kashida-PS) by integrating them together for an efficient formula. The proposed method provided a large hiding space (Al-Nofaie, S., Gutub, A. and Al-Ghamdi, M., 2021).

Alyousuf, F., and Roshidi, D. indicated the presentation of a set of concealment methods in the text and the measures of efficiency used in its evaluation, as it depended on the characteristics (Feature-Based Steganography) is the most used in most of the methods It is used to measure the efficiency of the methods of concealment in the text was according to the following percentages: security 34%, capacity 24%, durability 23%, and time included 19% (Alyousuf, F.Q.A. and Din, R., 2020).

Mustafa N. studied a new way to hide the text using invisible letters and compared the text with the hidden text. This research included the generation of a secret message in four main stages such as using the letter from the original message, choosing the appropriate cover text, dividing the text into blocks, and concealing the secret text. One of the advantages of this method is high secret production due to multiple levels of complexity to avoid the attacker (Mustafa, N.A.A., 2020).

Alanazi N. et al., stated a proposed method using (Unicode), in addition to using invisible and visual letters such as Kashida, ZWJs, ZWNJ, also using MMSPS to increase the ability of this method without reducing data safe. This research included the use of the method of contextual forms of Arabic letters to conceal in text. One of the advantages of this method is its high degree of safety in relation to other methods (Alanazi, N., Khan, E. and Gutub, A., 2022).

Sadie J. et al., presented a proposal on the basis of the color coding to conceal the text using the first two ways: depends on the substitution (Permutations) and the second: dependent on the numability system. One of the advantages of these two methods provide better concealment and high capacity in terms of ease of reading (Sadié, J.K. et al., 2020).

Khekan A. et al., dealt with a method using the points of Arabic letters and used the Arabic Semantic

Dictionary to hide many secret texts. Part of the Arabic language features were used to include the secret English message in the text cover to create the information of the information, where the secret text is converted to the bilateral system and then the T-5Be algorithm on it to reduce the size of the secret text by 37%. One of the advantages of this research is a high hidden resolution and cover storage capacity (Saibabu, P.C. and et al., 2019).

Akbar, F. introduced a new method using (Word-Shift) technology, where the research included providing one of the important factors, which is the level of safety in hiding information, as the message was converted into a series of including bits in the document file by converting distances between words (Idres, A.A. and Yaseen, H.I., 2023).

Thabit, R. et al., presented the color method and the coordinated voids (CSNTSTAGE) to solve the few capacity problems of hiding and not revealing the hidden text. The proposed method included the first two phases: the Huffman method is used to reduce the size of the secret text to be hidden and work to increase the number of bits that can be hidden in each location or letter of the original text. The second stage depends on the coloring of the color and the distance in the original text to solve the problem of the color difference between the colors of the cover and the resulting text after the concealment process (Thabit, R. et al., 2022).

Figueira, J. stated a way to hide information in the text using Winstein's Ideal Coding, which has the highest percentage of hiding information, where all possible synonyms can be used for words. Also, use Huffman coding to reduce the resulting text cover. One of the advantages of this method is one of the advantages of this method is a high hiding rate (Figueira, J., 2022).

Adeeb, O. and Kabudian, S. handled a way to hide secret data using artificial intelligence and long-term memory (LSTM) to increase a capacity of 45%. This research included the use of the Arabic language due to its large number of words, vocabulary, and linguistic meanings where the research relied on the previous Arabic poetry texts, where the linguistic accuracy was increased within the poetry formula with the use of an algorithm (Baudot code), where the secret data was hidden at the level of letters instead of words (Adeeb, O.F.A. and Kabudian, S.J., 2022).

Abdul Majeed, M. et al. presented a way to hide the information that combines encryption and pressing using multi-layer FPE coding and coding to reduce the size of secret data before hiding it. This research ensures the use of invisible unicode letters to include secret data in text files in English to produce STEGO files. One of the advantages of this method is a significant improvement in hiding information and lack of doubt about the cover file (Majeed, M.A. et al., 2022).

Osman B. et al. offered a method on the basis of color (RGB), red, green, blue, and the second is the theory of dividing the rest (SQRT) to hide in the text, where (RGB) was used on a scale of (0,0,0) to (15,15 , 15) To avoid suspicion of color, in addition to using the generation of random numbers (PRNG) to make secret messages dynamic with the creation of a table (homophonic) where a quantity of information is hidden by 77.4%. This research guarantees the hiding of information by changing the features of the letters (size, shape, style). One of the advantages of this method is a high secret ability (Osman, B. and et al., 2023).

Shakir, N. and Mahd, M. used a new, unaccounted method to improve the ability to include and increase the effect by collecting the grammar of the Arabic language and marks of formation (movements) with the Unicode method and discrimination in the use of concealment of information on the basis of using special letters (A – O- D – TH- R - Z) in the Arabic language in the Holy Quran. One of the advantages of this method is its high ability to hide and its high secrecy (Shakir, N.S. and Mahdi, M.S., 2023).

Khosravi, B. indicated a new way to hide in the text that depends on that there are different models (RGB, HSL, HSV) are used to represent color values, and that the very little difference between two different colors but close to the RGB system as the same coding in the system (HSL) The researcher relied on this method to do the concealment. The abundance of this method is a large treasury space for the invisible concealment process (Khosravi, B., 2023).

## 4 Proposed Method

This research uses the coding process to build a method of hiding in the text; most of the previous hiding methods depend on converting the secret message into ASCII code and then into binary form; after that, the obtained binary string is hidden by adopting one of the previous methods of hiding within the text. In our

proposed method, the most important ideas and methods that are adopted in building the proposed method will be identified as follows:

1. Search for the letters of the secret message within the cover text and encode the locations of those letters so that the code is two syllables (number, letter), the number represents the cover sequence word containing the character (of the secret message) and the letter represents the character sequence of the secret message within the word. An illustrative example of the coding process: Suppose we have the following:
Cover text:" the University of Mosul, College of Computer and Mathematic Sciences, Computer Science Department", and the secret message: "help Me" so the code sequence will be: "8D,1E,3E,6D,AA,3A,4E".

2. After obtaining a series of symbols (a combination of numbers and letters) that represent the locations of the secret message letters in the cover text, these symbols are converted into binary form and thus become ready for the embedding process.

3. The process of embedding: The previous ideas and methods used by researchers to conceal text were benefited from, and the following new methods for the concealment process were proposed:
A- If the cover word is of odd length and the binary number to be hidden is "0", then the first letter of the word is changed to a capital letter. If the binary number is "1", the first letter of the word is changed to a small letter.
B- If the cover word is of even length and the binary number to be hidden is "0", then the first letter of the word is changed to a small letter. If the binary number is "1", the first letter of the word is changed to a capital letter.

C- If the word is more than four letters long and its length is:
1- Even, the letter is changed before the middle. If the binary number is "0", we make it a lowercase letter, and if it is "1", we make it an uppercase letter.
2- Odd, then the middle letter is changed. If the binary number is "0", we make it a capital letter, and if the binary number is "1", we make it a lowercase letter.

D- If there are some punctuation symbols (. , , , ;), a space is added before them if the binary character is "0" or the space is removed if the binary character is "1".

E- At the end of each line of the cover text, a space is added if the binary character is "0" and a space is not added at the end of the line if the binary character is "1".

F- Retrieval process: In this process, we only have the stego-text. All words and passages that were used in the hiding process are searched analyzed, and the binary string is obtained. This string is converted according to the ASCII encoding, and the resulting ASCII string is then converted into a string of symbols. Finally, the resulting symbols' strings are compared with the stego-text, thus obtaining the characters of the secret message. Fig. 3 represent the proposed method flowchart.

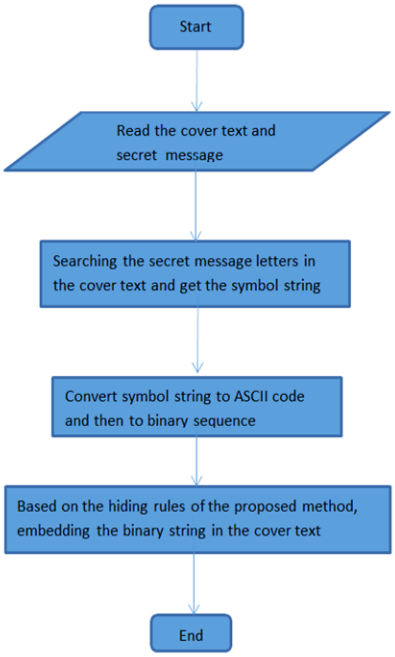


Figure 3. Embedding flowchart

The hidden text retrieval algorithm includes the following steps:
1- Read the carrier text according to the rules used in the hiding process, where:
A - If the word is of odd length and the first letter is large, this indicates that the value "0" is hidden, but if the first letter of the word is small, then the hidden value is "1".

B - If the word is of even length and the first letter is large, this indicates that the value "1" is hidden, but if the first letter of the word is small, then the hidden value is "0".

C- If the word is more than four letters long:

- Its length is even: if the letter before the middle is small, this indicates that the hidden value is "0", and if the letter before the middle is large, then the hidden value is "1".
- Its length is odd: If the letter before the middle is small, this indicates that the hidden value is "1", and if the letter is large, then the hidden value is "0".

D - If one of the punctuation symbols (.,,,, ;) comes and there is a space before it, then the hidden value is "0", and if there is no space, then the hidden value is "1".

E - At the end of each line of the carrying text, if there is a space, this indicates that the hidden value is "0", and if there is no space at the end of each line, then the hidden value is "1".

2- Convert the binary string into its corresponding letters and numbers according to the ASCII encoding, thus obtaining the symbol string.

3- By comparing the string of symbols with the carrier text, the secret message is obtained in its original form, complete and without any loss, including the spaces between the words and a number of lines, if any.

The retrieval algorithm can be illustrated with the flow chart in the Fig. 4.

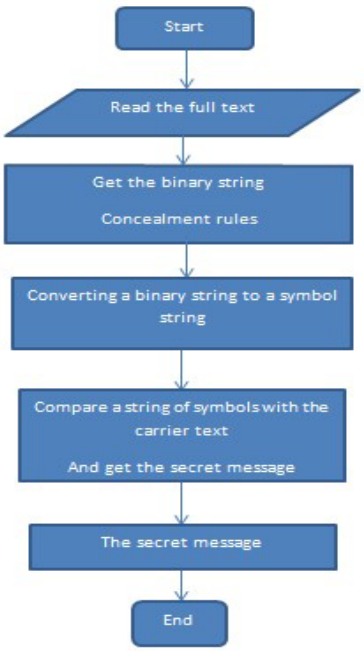


Figure 4. Extracting flowchart.

## 5. Analysis of the efficiency of the proposed method:

A. Most of the previous studies in the field of text hiding were based on the robustness criterion, so that the hidden text could be retrieved if it was opened using a different text editor on the receiving end. The proposed method successfully exceeded this criterion.

B. There are several metrics that can be used to measure the similarity between texts (cover text, text after hiding), including:

1. Edit Distance measures the number of operations (insertion, deletion, and replacement) required to convert one text to another.
2. Cosine Similarity coefficient is used to analyze the similarity of word patterns in texts.
3. The Jaccard Similarity scale measures the similarity between two groups by measuring the ratio of the common elements between them to the total elements.
4. The semantic similarity measure is used to measure the extent of similarity in meanings between texts.
5. Syntactic Structure Comparison focuses on the similarity or difference in the grammatical structure of sentences and phrases.

All of the above standards are successfully exceeded in the proposed method, as it relies on changing the case of letters only to perform the masking process without deletion or addition.

C. During the practical implementation of the proposed method, the length of the binary string was calculated, as well as the number of words used for hiding, in addition to the number of letters that were changed, as we noticed that the number of words used to perform the hiding process is less than the number of binary numbers to be hidden since there are many words that carry specifications. A certain number can be used to hide more than one binary number, and since when changing the case of a specific letter (uppercase or lowercase) depending on the binary character to be hidden, the case of the letter may be similar in origin to the case to which it is requested to be changed, and therefore no change will occur to the case of that letter, which increases the efficiency of the method, as shown in Table 1.

Table 1. Letters needed for concealment.

| The length of the binary message | The number of words | Number of characters |
|---|---|---|
| 152 | 97 | 81 |
| 232 | 147 | 97 |
| 128 | 80 | 63 |
| 120 | 76 | 55 |

## 6. Conclusion

A number of challenges arise when using text files as a cover for information concealment. The most significant ones are the restricted amount of space available for the concealment process, the high sensitivity of any modifications made to the text file, and the possibility that using a different text editor will cause the concealment process to fail. Text files are no longer used as a means of information concealment as a result of all these limitations. In this research, a new method of text steganography was presented that includes two levels. The first is the encoding process, which involves searching for the letters of the secret message within the cover text and obtaining the symbol string. The second level involves converting the symbol string to binary form and hiding it in the cover text using suggested rules.

The use of two levels of concealment in the proposed method provides a high level of protection for the hidden text and thus makes it difficult for the attacker to discover the secret message.

## References

Adeeb, O.F.A. and Kabudian, S.J., 2022. Arabic text steganography based on deep learning methods. IEEE Access, 10, pp.94403-94416.

Agarwal, M., 2013. Text steganographic approaches: a comparison. arXiv preprint arXiv:1302.2718.

Akbar, F.C., Purboyo, T.W. and Latuconsina, R., 2020. A study of text steganography methods. Journal of Engineering and Applied Sciences, 15(2), pp.369-372.

Alanazi, N., Khan, E. and Gutub, A., 2022. Inclusion of unicode standard seamless characters to expand Arabic text steganography for secure individual uses. Journal of King Saud University-Computer and Information Sciences, 34(4), pp.1343-1356.

Al-Nofaie, S., Gutub, A. and Al-Ghamdi, M., 2021. Enhancing Arabic text steganography for personal usage utilizing pseudo-spaces. Journal of King Saud University-Computer and Information Sciences, 33(8), pp.963-974.

Alyousuf, F.Q.A. and Din, R., 2020. Analysis review on feature-based and word-rule-based techniques in text steganography. Bulletin of Electrical Engineering and Informatics, 9(2), pp.764-770.

Chaw, A.A., 2019. Text steganography in Letter of Credit (LC) using synonym substitution based algorithm. International Journal for Advance Research and Development, 4(8), pp.59-63.

Figueira, J., 2022. A Survey on Semantic Steganography Systems. arXiv preprint arXiv:2203.12425.

Hamdan, A.M. and Hamarsheh, A., 2016. AH4S: an algorithm of text in text steganography using the structure of omega network. Security and Communication Networks, 9(18), pp.6004-6016.

Idres, A.A. and Yaseen, H.I., 2023. Text Steganography Techniques: A Review. International Research Journal of Innovations in Engineering and Technology, 7(11), p.648.

Khan, Y., Algarni, A., Fayomi, A. and Almarashi, A.M., 2021. Disbursal of text steganography in the space of double-secure algorithm. Mathematical Problems in Engineering, 2021, pp.1-9.

Khosravi, B., 2023. Text steganography by changing the black color. AUT Journal of Mathematics and Computing.

Majeed, M.A., Sulaiman, R. and Shukur, Z., 2022. New Text Steganography Technique based on Multilayer Encoding with Format-Preserving Encryption and Huffman Coding. International Journal of Advanced Computer Science and Applications, 13(12).

Majeed, M.A., Sulaiman, R., Shukur, Z. and Hasan, M.K., 2021. A review on text steganography techniques. Mathematics, 9(21), p.2829.

Mandal, K.K. and Singh, P.K., 2019, March. Information Hiding in Text Steganography: A Different Approach. In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE).

Mustafa, N.A.A., 2020. Text hiding in text using invisible character. International Journal of Electrical and Computer Engineering, 10(4), p.3550.

Osman, B., Yahya, N.I., Mohd Zaini, K. and Abdullah, A., 2023. Text Steganography Using the Second Quotient Remainder Theorem and Dark Colour Schemes. Journal of Computational Innovation and Analytics (JCIA), 2(1), pp.21-40.

Sadié, J.K., Metcheka, L.M. and Ndoundam, R., 2020.

Two high capacity text steganography schemes based on color coding. arXiv preprint arXiv:2004.00948.

Saibabu, P.C., Sai, H., Yadav, S. and Srinivasan, C.R., 2019. Synthesis of model predictive controller for an identified model of MIMO process. Indonesian Journal of Electrical Engineering and Computer Science, 17(2), pp.941-949.

Shakir, N.S. and Mahdi, M.S., 2023. Using special letters and diacritics in Steganography in holy Quran. Iraqi Journal for Computers and Informatics, 49(2), pp.1-8.

Sun, B., Li, Y., Zhang, J., Xu, H., Ma, X. and Xia, P., 2023. Topic Controlled Steganography via Graph-to-Text Generation. CMES-Computer Modeling in Engineering & Sciences, 136(1).

Taleby Ahvanooey, M., Li, Q., Hou, J., Rajput, A.R. and Chen, Y., 2019. Modern text hiding, text steganalysis, and applications: a comparative analysis. Entropy, 21(4), p.355.

Thabit, R., Udzir, N.I., Yasin, S.M., Asmawi, A. and Gutub, A.A.A., 2022. CSNTSteg: Color spacing normalization text steganography model to improve capacity and invisibility of hidden data. IEEE Access, 10, pp.65439-65458.

## Biography

**Alaa Abdullah Idress,** he is a high diploma (postgraduate) student at the Department of Computer Science, College of Computer Science and Mathematics, University of Mosul. He is interested in the research related to the security including steganography, cyphering, and cryptography.

**Yaseen Hikmat Ismael,** he is an associate professor at the Department of Computer Science, College of Computer Science and Mathematics, University of Mosul. His Ph.D. was in the security field. He is interested in the topics related to security field and the integration with other fields in computer science.