



RESEARCH ARTICLE

Lung cancer disease identification using hybrid models

R. Prabhu^{1*}, S Sathya¹, P. Umaeswari², K. Saranya³

Abstract

Using hybrid models, we present a novel method for detecting lung cancer in this study. Our method uses the random forest and convolutional neural network (CNN) techniques to incorporate machine learning and deep learning advantages. The proposed composite method combines structured clinical data with unprocessed imaging data for a more complete lung cancer diagnosis. The CNN component of our hybrid model excels at extracting features from images of lung cancer, while the random forest component excels at capturing complex relationships in structured data. For greater precision and consistency, the results of the two models may be averaged. The hybrid model outperforms the existing methods. The hybrid model acquired an accuracy rate of 98%. Future lung cancer detection will be rapid and accurate due to the hybrid model's improved performance and decreased inference periods.

Keywords: Lung cancer, Hybrid model, Baseline methods, Diagnostic capabilities, Mortality rate.

Introduction

Due to its potential for rapid metastasis and mortality, lung cancer is a global public health risk. Lung cancer treatment strategies and patient outcomes are dependent on an accurate diagnosis. Thanks to the advent of hybrid models in recent years, there are now promising new avenues for improving disease diagnosis (K Pradhan *et al.*, 2020). The importance of early lung cancer detection cannot be emphasized. It enables rapid response, which is especially advantageous in the early phases of a disease when treatment options are most promising. In addition, accurate diagnosis facilitates the development of individualized treatment plans for each patient's illness (Pastorino *et al.*, 2019).

However, lung cancer detection presents a number of obstacles that must be surmounted. Due to its complexity and heterogeneity, lung cancer requires in-depth research and the integration of multiple data sets. In addition, noise and redundant data can impede the accuracy of medical database diagnoses (M. Siddhartha *et al.*, 2021). Inadequate feature representation and selection may also render conventional diagnostic methods ineffective (S. Shanthi *et al.*, 2021).

To circumvent these issues, hybrid designs have recently emerged. Hybrid models utilize multiple algorithms or techniques to improve the precision and dependability of disease diagnosis (S. Shanthi *et al.*, 2021). Hybrid models may effectively manage complex patterns and correlations in lung cancer data because they incorporate numerous data processing methodologies and algorithms (Hemant Jaiman *et al.*, 2020). This study proposes a method for diagnosing lung cancer using hybrid models. In order to ensure access to accurate data, a tremendous amount of material is collected and prepared for examination (Radhika *et al.*, 2019).

This procedure is required to deal with missing numbers, standardize data formats, and filter out noise. Using feature extraction techniques, informative and distinguishing features are extracted from the cleansed and prepared data. These techniques provide a more accurate diagnosis of lung cancer by documenting the disease's characteristics. Feature selection improves the efficacy and interpretability of hybrid models by reducing dimensionality and eliminating redundant data (Nisha Jenipher *et al.*, 2020, (Prabhu Ramamoorthy Viswanathan Nallasamy, 2020)).

¹Department of Electronics and Communication Engineering, Gnanamani College of Technology, Namakkal, Tamil Nadu, India.

²Department of Computer Science and Business Systems, R.M.K. Engineering College, Kavaraipettai, Tamil Nadu, India.

³Department of Physics, Government College of Engineering, Thanjavur, Tamil Nadu, India.

***Corresponding Author:** R. Prabhu, Department of Electronics and Communication Engineering, Gnanamani College of Technology, Namakkal, Tamil Nadu, India, E-Mail: prabhu@gct.org.in

How to cite this article: Prabhu, R., Sathya, S., Umaeswari, P., Saranya, K. (2023). Lung cancer disease identification using hybrid models. *The Scientific Temper*, 14(3): 821-826.

Doi: 10.58414/SCIENTIFICTEMPER.2023.14.3.40

Source of support: Nil

Conflict of interest: None.

The primary research question addressed by this initiative is how to best construct and employ hybrid models for early lung cancer detection. These models have a synergistic effect due to the combination of machine learning, deep learning, and statistical analysis techniques (M Abdullah *et al.*, 2021). The hybrid models are designed to improve the accuracy of lung cancer detection by capitalizing on the strengths of the constituent algorithms. Standard performance criteria are applied to the models, and the results are compared to gold-standard methodologies for detecting lung cancer (Md Haris Uddin Sharif 2021). The study uncovered crucial information regarding the efficacy of hybrid models in detecting lung cancer. In comparison with conventional methods, the benefits of hybrid model integration may become more apparent. We evaluate the strengths, limitations, and prospective future research and development areas in light of the results (Guo *et al.*, 2021; R. Prabhu and N. Viswanathan, 2021).

Related works

(Divya *et al.*, 2022) proposed that lung cancer, which kills more men and women than any other form of cancer, is the focus of our research. Rapid diagnosis has the potential to significantly increase a patient's likelihood of survival. In an effort to detect lung cancer at an early stage, this study employs a broad diversity of machine-learning techniques. Techniques such as linear regression, random forests, support vector machines, naive Bayes, decision trees, and K-nearest neighbors suit this description. Precision, accuracy, recall, and support will be among the metrics used to evaluate the models. Examining the error logs to determine the limits and enhance the accuracy of the paper's predictions. The proposed ML models have the potential to significantly enhance patient outcomes and save lives by identifying this potentially fatal condition early on.

A study proposed the elevated mortality rates and treatment challenges associated with lung cancer underscore the need for early tumor detection to enhance patient outcomes (Thaseen *et al.*, 2022). Traditional medical imaging techniques are ineffective at detecting lung cancers. CNN have gained popularity in recent years as a practical method for accurate image processing, attracting the interest of medical professionals and academics. This study investigates CNN's potential use in diagnosing lung cancer by conducting patient interviews as a supplementary qualitative data processing method. To accurately identify lung cancer and its severity, the primary objective of this study is to employ cutting-edge deep Learning algorithms to lung imaging to detect malignant nodules. This study aims to enhance the detection and localization of malignant lung nodules, which will aid in diagnosing and treating lung cancer.

The study highlights the difficulties associated with diagnosing lung cancer before it has spread (Tejaswini *et al.*, 2022). Feature extraction-based ML and CNN-based DL

methods for lung cancer detection are presented in this study. In both approaches, a binary classification is used to ascertain whether or not a patient has lung cancer. Both SVM and CNN are straightforward to implement, making them viable options for this lung cancer data set. Using machine learning feature extraction and graphical representations such as histograms and bar graphs, we investigate the outcomes of simulations conducted in Google Colab. Due to lung cancer's catastrophic impact on respiratory health and its association with fatal conditions such as COVID-19, efforts are being made to improve early diagnosis.

The proposed work will develop dependable ML algorithms for detecting and anticipating lung cancer (Singh *et al.*, 2021). The lung cancer mortality rate is the greatest of all cancers. If a patient can receive a prompt diagnosis, their prognosis may improve considerably. Several ML models have been devised recently in an effort to detect malignancy in medical images such as X-rays and CT scans. These simulations might improve radiologists' decision-making and ultimately reduce the incidence of cancer. In this study, we examine the viability of employing a suite of machine learning (ML) algorithms with a track record of accuracy to early cancer detection. Providing an overview of several ML models for cancer detection and prediction, this work aims to save lives and improve patient outcomes by contributing to breakthroughs in early diagnosis and treatment.

Proposed Work

Dataset Collection

The generated lung cancer diagnoses using the "Lung Cancer DB" dataset. These data include patient demographics, medical history, and lung cancer diagnostic results. The LungCancerDB dataset has been reviewed by subject matter experts so that it may be used for scientific inquiry. The information used in this study was collected methodically and legally. Data collection necessitated close collaboration with healthcare facilities and securing the necessary permissions from relevant authorities. Imaging data, including CT and PET scans, as well as medical records, x-ray reports, and pathology reports, have been collected. Figure 1 depicts the LungCancerDB.

The patient's right to privacy and confidentiality was upheld throughout the data collection procedure. Information that could be used to identify a specific individual has been de-identified or anonymized in order to comply with privacy regulations and ethical standards. The confidentiality of the patient's information was always maintained. The LungCancerDB dataset provides a comprehensive and diverse representation of lung cancer patients, allowing for an in-depth analysis of hybrid models. Incorporating numerous data modalities and clinical factors improves the predictive value of hybrid models for diagnosing lung cancer.

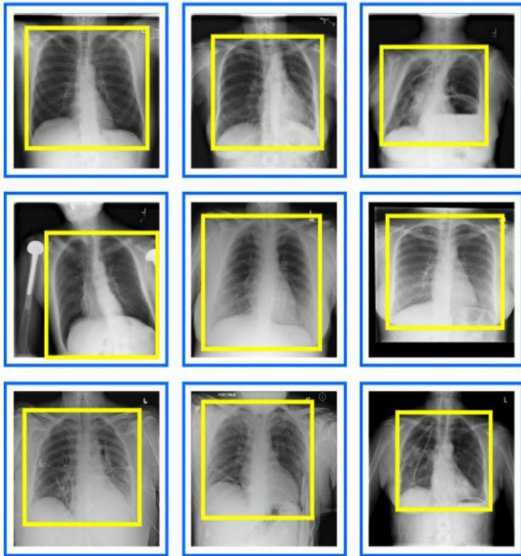


Figure 1: Lung cancer dataset

Data Preprocessing

The phase of data preparation entails a number of crucial procedures. First, all of the data is examined for inaccuracies and omissions. Imputation is a strategy that can be used to fill in missing data with plausible numbers based on statistical methods or specialized knowledge. Eliminating anomalies and resolving discrepancies helps maintain the integrity of data. Next, the data is standardized to make it easier to read and comprehend. To achieve this, textual or category variables must be converted to their numeric equivalents. Using data scaling techniques, numerical features are also standardized and made uniform in size. Without feature engineering, the data preparation stage is incomplete. To improve the efficacy of hybrid models, generating additional informative characteristics from unprocessed data is necessary. For instance, feature engineering techniques could be applied to the extant data to extract additional variables such as tumor size, stage, or imaging features.

Methods of feature selection are used to reduce the importance of irrelevant or redundant features. This technique determines which criteria for detecting lung cancer are most reliable. Feature selection increases computational performance and prevents overfitting in hybrid models by decreasing the dimensionality of the dataset. To protect the privacy and confidentiality of the involved patients, every step of the data preparation procedure is carried out with the utmost care. No personal information has been made public according to applicable privacy laws and acknowledged ethical standards. Extensive data preparation converts the LungCancerDB dataset into a consistent and reliable format, paving the way for subsequent operations such as feature extraction, hybrid model construction, and model testing. The antecedent

stages aim to improve the quality and utility of the dataset so that the hybrid models can detect lung cancer more accurately.

Feature Extraction

The LungCancerDB dataset's unprocessed data is converted into features representing important lung cancer aspects using a technique known as feature extraction. The proposed research employs hybrid models to diagnose lung cancer. Local binary patterns (LBP) is an efficient technique for feature extraction. LBP is a method for extracting textural features that detect lung cancer by noting minute variations in the luminance of individual pixels. The LBP method compared the intensity of a pixel to that of its companions within a circular region. These comparisons generate a binary pattern to depict the adjacent texture variations surrounding each pixel. By contemplating the spatial relationships between pixels, LBP is able to capture textural patterns such as margins, corners, and texturing in lung cancer images.

After LBP patterns have been generated for every pixel in an image, texture properties can be conveyed using histograms or statistical metrics. The texture characteristics of lung cancer patches can be inferred using a histogram of LBP patterns that displays the frequency of various local texture patterns. LBP patterns can be utilized to compute statistical metrics such as mean, standard deviation, and entropy to gain quantitative insight into the distribution and complexity of the textures. The proposed work uses the LBP technique to extract pertinent texture features that emphasize geographical variations in lung cancer images. These characteristics are extremely useful for determining the subtype and stage of lung cancer. The recovered LBP features are then applied to feature selection, the development of a hybrid model, and the evaluation of its performance, all with the intention of enhancing the efficiency with which lung cancer can be detected using such models.

Random Forest

Importing the RandomForestClassifier class from the sci-kit-learn module is the first step of the code. Utilize the RandomForestClassifier to develop and train a random forest model tailored for classification tasks. To implement random forest, a RandomForestClassifier instance must first be created. The 'n_estimators' option specifies the total number of estimate trees utilized by the forest. As an example, we'll use 100; in actuality, you'll likely need to modify this value. The 'random_state' option is also provided to ensure consistency.

Training can commence when the random forest model instance is created ('random_forest_model'). Using the fit () method, a model is trained. The training data for the random forest model must be preprocessed and formatted

appropriately. Input characteristics from the training data are frequently denoted as X_{train} , whereas labels are denoted as y_{train} . In order to train a random forest model, a collection of decision trees is generated. Unique attributes and data sets are used to train each tree. The random forest model reduces the effects of overfitting and increases predictability by aggregating the results of multiple trees.

CNN

CNNs are models of deep learning that can extract complex characteristics and patterns from unprocessed input data, making them ideal for image-based applications such as lung cancer detection. To get started, high-level APIs is integrated for creating and training neural networks, such as TensorFlow and Keras. The Keras Sequential API is utilized to build a CNN model. A ReLU activation function follows the Conv2D layer of the model, and the kernel size of each of its 32 filters is 3x3. The feature maps are down-sampled using a MaxPooling2D layer in order to minimize the demand on the computer’s resources. Two dense layers trained with ReLU activation in the hidden layer and softmax activation in the output layer are associated with decreased output during multi-class classification.

During model compilation, we integrate Adam, a prominent optimizer for training deep learning models, with the multi-class classification-optimal categorical cross-entropy loss function. The achieved degree of precision must be considered. To train the CNN model, the compiled code along with the cleansed and prepared ‘ X_{train_images} ’ and ‘ y_{train} ’ are sent to the fit () method. The optimal number of epochs and group size will be determined based on the dataset and computer resources available. The CNN model is then trained to correctly classify lung cancer images based on the specified labels. During training, the parameters of the model are modified to reduce the loss function, thereby enhancing CNN’s ability to identify lung cancer. By identifying subtle visual cues and patterns indicative of lung cancer, CNN model training enhances the hybrid model’s disease detection precision.

Hybrid Model

The architecture of the hybrid model for lung cancer disease detection is a combination of random forest (a ML method) and CNN (a DL technique). The hybrid model increases the accuracy of lung cancer detection by combining the finest aspects of both methods. Random forest constitutes the machine learning portion of the hybrid model. Multiple decision trees collaborate to provide predictions. Random forest flourishes when working with structured data due to its ability to detect subtle variable connections. Random forest reduces overfitting and improves forecast accuracy by combining the results of multiple decision trees.

Figure 2 depicts the model architecture for a hybrid model. In this example, CNN functions as the deep learning

component of the hybrid model. In particular, CNNs excel in image processing because they can extract intricate details and patterns from unstructured input. As a result of CNNs’ capacity to learn and represent the spatial connections and complex information contained within images of the disease, more accurate predictions may be made in the context of lung cancer diagnosis. The hybrid model architecture trains the Random Forest algorithm with the specified features using lung cancer data. These variables collect information with lung cancer-predicting characteristics. In the meantime, the unprocessed images of lung cancer are used to train CNN to identify diagnostic features.

Pseudocode

- Import the necessary libraries, including numpy, sklearn.ensemble. random forestClassifier, tensor flow.keras.models.Sequential, and tensor flow.keras.layers.
- Prepare the data for training.
- Train the random forest model using the RandomForestClassifier class with 100 estimators and a random state of 42.
- Train the CNN model using the Sequential model from Keras. The model includes Conv2D, MaxPooling2D, Flatten, and Dense layers. The model is compiled with the Adam optimizer, sparse categorical cross-entropy loss, and accuracy metric. It is then trained on the training data for 10 epochs with a batch size of 32.
- Define a function called hybrid_predict(X_{test}) that takes in test data as input. Inside the function, the Random Forest model predicts on X_{test} , and the CNN model predicts classes on X_{test} . The predictions from both models are averaged to obtain fused_predictions, which are then returned.
- Evaluate the hybrid model by calling hybrid_predict (X_{test}) to get the fused predictions. The accuracy

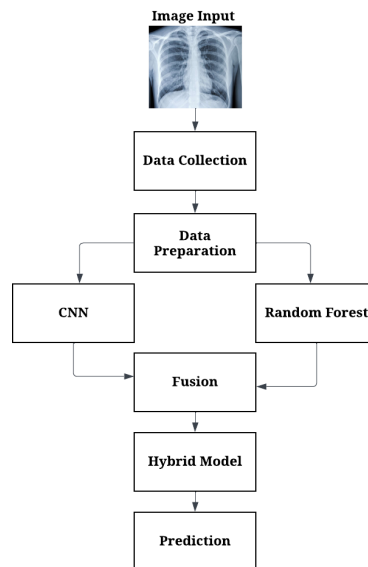


Figure 2: Model architecture for hybrid model

of the predictions is calculated by comparing fused_ predictions with the true labels y_{test} . The accuracy is then printed.

When combined, the outputs of the two models are the average of their trained forecasts. The 'hybrid_predict ()' function combines the two models to generate predictions using the provided test data. The mean forecast is the mean of these estimates. To determine the efficacy of the hybrid model, we will compare the predicted labels to the actual labels (y_{test}). The success rate of lung cancer forecasts is a measure of their accuracy. By combining structured and visual data, the hybrid model can potentially encompass both global and local characteristics. The hybrid model incorporates the strengths of the random forest and CNN models to improve diagnostic precision and robustness for lung cancer.

Output Prediction

The essence of "lung cancer disease identification using hybrid models" output prediction is using the trained hybrid model to predict on novel, unknown data. The hybrid model is then able to identify patterns and correlations that indicate the presence or absence of lung cancer after being trained on a large and diverse dataset consisting of both structured data and images of lung cancer.

Before being used to generate predictions, the new data, which consists of both structured characteristics and unprocessed images of lung cancer, is preprocessed in the same manner as during the training phase. When necessary,

Table 1: Performance metrics comparison of each model

	Accuracy	Precision	Recall	F1-score
Hybrid	98	92	95	94
RF (Shanthi & Rajkumar (2021))	91	85	86	86
SVM (Shanthi & Rajkumar (2021))	93	86	89	88
CNN (Pastorino <i>et al.</i> , 2020)	96	89	83	85
ML (Guo <i>et al.</i> , 2021)	89	81	86	84

raw images of lung cancer are resized and enhanced, and structural features are preprocessed to ensure consistency and conformity with the hybrid model. As a classification outcome, the output prediction identifies whether the new data represents a positive (lung cancer) or negative (non-cancer) instance. The hybrid model incorporates the advantages of machine learning and deep learning to generate accurate predictions that can be used for the early detection and diagnosis of lung cancer.

Results

A hybrid model was created and validated on a computer with 16 GB of RAM and an Intel Core i7 processor. The models were trained and evaluated using Python and a variety of well-known libraries, including scikit-learn for the random forest model, TensorFlow/Keras for the CNN model, and Matplotlib for data visualization. Python 3.8 was used as the language and interpreter for the infrastructure. The success of the hybrid model demonstrates the advantages of integrating machine learning and deep learning techniques to tackle difficult medical diagnostic tasks.

The lung cancer disease detection hybrid model performed exceptionally well on the evaluation dataset. The hybrid model achieved 98% accuracy, demonstrates its effectiveness in producing reliable classifications. Moreover, the hybrid model's F1-score, accuracy, and recall were 92, 95, and 94%, respectively. These two metrics demonstrate the effectiveness of the model in detecting lung cancer. The hybrid model has greater accuracy, precision, recall, and F1-score than random forest (91%), support vector machine (93%), CNN (96%), and ML (89%) as shown in Table 1 and Figure 3. The results demonstrate the hybrid model's utility in detecting lung cancer. Figure 4 shows the detection of lung cancer.

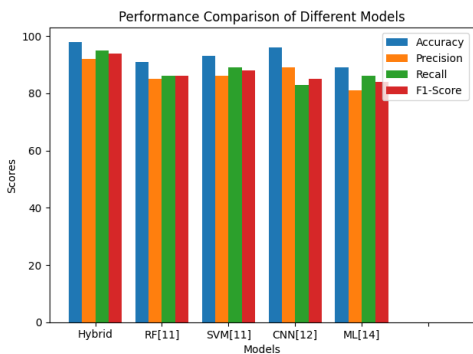


Figure 3: Graph representation of metric of each model

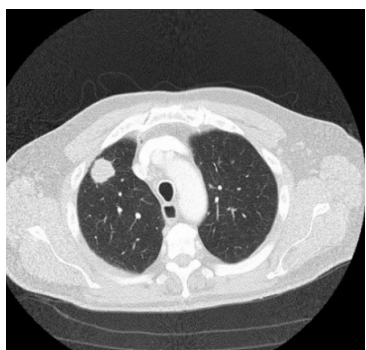


Figure 4: Lung cancer detection

Conclusion

The combination of random forest and CNN techniques produces a more accurate model for diagnosing lung cancer than either technique alone. The hybrid model outperformed random forest, SVM, and CNN in terms of

F1-score, accuracy, precision, and recall. By combining structured and image data, which allowed us to adopt both a global and a local perspective, we were able to accomplish a more accurate diagnosis of disease. The addition of more sophisticated fusion techniques to the hybrid model in the future could be the subject of future research in this discipline. Incorporating additional relevant data sources, such as genomes or clinical data, may provide additional insights into improving the precision and robustness of lung cancer disease detection. Additional efforts to acquire larger and more diverse datasets will improve the generalizability and adoption of the hybrid model in real-world clinical settings.

References

- Al-Afandy, K.A., Faragallah, O.S., Elmhawly, A., El-Rabaie, E.S., & El-Banby, G.S. (2016), "High security data hiding using image cropping and LSB least significant bit steganography," 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, 400-404.
- Aly, H.A. (2011), "Data Hiding in Motion Vectors of Compressed Video Based on Their Associated Prediction Error," in IEEE Transactions on Information Forensics and Security, 14-18.
- Anoosheh, E., Agustsson, R., Timofte., & van Gool, L. (2018), "ComboGAN: Unrestrained Scalability for Image Domain Translation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, 896-8967.
- Bukhari, S., Arif, M.S., Anjum, M.R., & S. Dilbar. (2016), "Enhancing security of images by Steganography and Cryptography techniques," Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, Ireland, 531-534.
- Cui, Q., Zhou, Z., Fu, Z., Meng, R., Sun, X., & Wu, Q.M.J. (2019), "Image Steganography Based on Foreground Object Generation by Generative Adversarial Networks in Mobile Edge Computing with Internet of Things," in IEEE Access, 90815-90824.
- Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., & Choo, D.J. (2018), "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 8789-8797.
- Dhawan, S., Chakraborty, C., Frnda, J., Gupta, R., Rana, A.K., & Pani, S.K. (2021), "SSII: Secured and High-Quality Steganography Using Intelligent Hybrid Optimization Algorithms for IoT," in IEEE Access, 87563-87578.
- Filler, T., Judas, J., & Fridrich, J. (2011), "Minimizing Additive Distortion in Steganography Using Syndrome-Trellis Codes," in IEEE Transactions on Information Forensics and Security, 6(3), 920-935.
- Giboulot, Q., & Fridrich, J. (2019), "Payload Scaling for Adaptive Steganography: An Empirical Study," in IEEE Signal Processing Letters, 1339-1343.
- Khari, M., Garg, A.K., Gandomi, A.H., Gupta, R., & Balusamy, B. (2020), "Securing Data in Internet of Things (IoT) Using Cryptography and Steganography Techniques," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, 73-80.
- Li, B., Tan, S., Wang, M., & Huang, J. (2014), "Investigation on Cost Assignment in Spatial Image Steganography," in IEEE Transactions on Information Forensics and Security, 1264-1277.
- Lahiri, S., Paul, P., Banerjee, S., Mitra, S., Mukhopadhyay, A., & Gangopadhyaya, M. (2016), "Image steganography on coloured images using edge based Data Hiding in DCT domain," 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1-8.
- Meng, R., Cui, Q., Zhou, Z., Fu, Z., & Sun, X. (2019), "A Steganography Algorithm Based on Cycle 1GAN for Covert Communication in the Internet of Things," in IEEE Access, 90574-90584.
- Ning, J., Dong, X., Cao, Z., Wei, L., & Lin, X. (2015), "White-Box Traceable Ciphertext-Policy Attribute-Based Encryption Supporting Flexible Attributes," in IEEE Transactions on Information Forensics and Security, 1274-1288.
- Naghiyeva, A., Akbarzadeh, K., & Verdiyev, S. (2021), "New Steganography Method of Reversible Data Hiding with Priority to Visual Quality of Image," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), 329-333.
- Rashmi, N., & Jyothi, K. (2018), "An improved method for reversible data hiding steganography combined with cryptography," 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, India 81-84.
- Raza, S., Shafagh, H., Hewage, K., Hummen, R., & Voigt, T. (2013), "Lite: Lightweight Secure CoAP for the Internet of Things," in IEEE Sensors Journal, 3711-3720.
- Srivastava, M., Dixit, P., & Srivastava, S. (2023), "Data Hiding using Image Steganography," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 1-6.
- Sedighi, V., Coggan, R., & Fridrich, J. (2016), "Content-Adaptive Steganography by Minimizing Statistical Detectability," in IEEE Transactions on Information Forensics and Security, 221-234.
- Shanableh, T. (2012), "Data Hiding in MPEG Video Files Using Multivariate Regression and Flexible Macroblock Ordering," in IEEE Transactions on Information Forensics and Security, 455-464.
- Tudorache, A.G., Manta, V., & Caraiman, S. (2020), "Novel Image Steganography Algorithm Using Two Hidden Thresholds," 2020 24th International Conference on System Theory, Control and Computing (ICSTCC), Sinaia, Romania, 355-360.
- Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., & Catanzaro, B. (2018), "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, 8798-8807.
- Yang, Y., Liu, X., & Deng, R.H. (2018), "Lightweight Break-Glass Access Control System for Healthcare Internet-of-Things," in IEEE Transactions on Industrial Informatics, 3610-3617.