BBRC
Bioscience Biotechnology
Research Communications

Improved Flower Pollination Algorithm for the Detection of Lung Cancer Detection in Humans Text Based Mining

Akey Sungheetha[1], Vankayalapati. Sahiti[2], Madiajagan M[3], R. C. Narayanan[4], S. Selvakanmani[5] and T. Ch. Anil Kumar[6]

[1]*Department of CSE, SoEEC, Adama Science and Technology University, Shewa, Ethiopia.*

[2]*Department of Electronics and Communication, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (DT), Andhra Pradesh, India.*

[3]*School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.*

[4]*Department of CSE, Sona College of Technology, Salem, Tamil Nadu, India.*

[5]*Department of Computer Science and Engineering, Velammal Institute of Technology, Velammal Knowledge Park, Chennai, Tamil Nadu, India.*

[6]*Department of Mechanical Engineering, Vignan's Foundation for Science Technology and Research, Vadlamudi, Guntur, Andhra Pradesh, India.*

Corresponding author email: sun29it@gmail.com

ABSTRACT

Lung cancer is a worldwide threat to humanity since its cells grow uncontrollably inside lungs leading to increased mortality rate. The lung cancer often poses serious breathing issues and it is been contributed majorly by smoking and inhaling smoke. Even with high medical advancements, the effective treatment and curing of lung cancer are not effective till date. Proper precautions and earlier stage detection may reduce the cancer to spread the entire organ. It is hence necessary to check with least minimal data i.e. text data can provide even a greater effectiveness in diagnosing the lung condition. Meta-heuristic algorithm influences greatly with its computational capability and offers stronger prediction of lung cancer at earlier stage with accurate analysis. In this paper, we develop a classification system using flower pollination algorithm (FPA) that tends to classify the medical text documents. The FPA algorithm classifies the medical text documents to diagnose the lung cancer in humans. The FPA is applied as an intelligent algorithm that imitates the behavior of pollination in flowering plants to identify the essential classes of lung cancer. It finds the relationship between the pollens to identify the essential classes based on flower position. The simulation is conducted to validate the effectiveness of the model with other meta-heuristic optimization methods that include bee colony optimization, and ant colony optimization algorithm. The results of simulation show that the proposed method undergoes effective classes of lung cancer than other existing methods that includes accuracy, sensitivity, specificity, f-measure and mean average percentage error.

KEY WORDS: CLASSIFICATION, LUNG CANCER, FLOWER POLLINATION ALGORITHM.

Biosc Biotech Res Comm P-ISSN: 0974-6455 E-ISSN: 2321-4007

Identifiers and Pagination	**Article Information**
Year: 2021 Vol: 14 No (7) Special Issue	Received: 11th May 2021
Pages: 229-234	Accepted after revision: 18th July 2021

DOI: *http://dx.doi.org/10.21786/bbrc/14.7.52*

INTRODUCTION

Lung cancer is the deadliest illness and the leading cause of death in the modern world. Lung cancer has a larger impact on humans, and it is predicted that it will now rank seventh in the death rate index, accounting for 1.5% of global mortality. Lung cancer starts in the lungs and progresses to the brain. Despite this, the lung cancer detection procedure is very ineffective because doctors will only be able to detect the illness after it has progressed to an advanced stage (Metovic, et al 2021, Abdullah and Abdulazeez, 2021, Malika, and Jaina, Yuvaraj, et al 2021).

As a result, early detection before the final stage is critical to lowering the mortality rate by successful monitoring. The survival rate is very promising even after proper medication and diagnosis. Lung cancer survival rates vary from person to person. Age, gender, race, and health status all play a role. Meta-heuristic is becoming increasingly important in the diagnosis and prediction in the early stages of human life (Kerr, K.M., et al 2021, Lai-Kwon, J., et al 2021, Natarajan, Y., et al 2021 and Miller, H.A., et al 2021). Meta-heuristic simplifies and predicts the diagnosis process. Meta-heuristic has also dominated the medical industry in recent years. Meta-heuristic models are currently being used in the county's health-care industry. Meta-heuristic may be used to investigate the actual diagnosis of diseases.

The meta-heuristic approach facilitates data analysis and processing of real characteristics or facts, allowing for the identification of disease problem creators. It aids medical professionals in determining the root cause of diseases. Picture processing: Image recognition has been shown to be reliable and useful through different Meta-heuristic processes. This enables the concerned doctors to make a more accurate diagnosis of the diseases, saving money and time while increasing the benefit proportion (Gowrishankar, J., et al 2020, Shaikh, S., et al 2021, Agazzi, G.M., et al 2021, Debata, P.P., et al 2021, Senthilkumar, P., 2021, Moser, S.S., et al 2021, Masud, M., et al 2021, and Liu, R., Rizzo, S., et al 2021).

As a result, the challenge has been solved, and the drug industry will now use the meta-heuristic method for processing. Meta-heuristic aids in the prediction of disease severity and outcome. The use of a meta-heuristic approach to monitor epidemic outbreaks allows for early detection and intervention. Meta-heuristic applications must also be streamlined in order to become more standardised and accurate (Masquelin, A. H., et al 2021, Triplette, M., et al 2021, Bright, R., et al 2021, Ma, X., et al 2021, Alzu'bi, A., et al 2021, and Montelongo González, E. E., et al 2020). As a result, further advancements in meta-heuristic algorithms would aid doctors and wellness catalysts in making accurate clinical decisions with high reliability and precision (Yuvaraj, N.,et al 2021, Saravanan V, et al 2016 and Saravanan V, et al 2016).

In this paper, we develop a classification system using flower pollination algorithm (FPA) that tends to classify the medical text documents. The FPA algorithm classifies the medical text documents to diagnose the lung cancer in humans. The FPA is applied as an intelligent algorithm that imitates the behavior of pollination in flowering plants to identify the essential classes of lung cancer. It finds the relationship between the pollens to identify the essential classes based on flower position. Related work : González et al. 2020 developed a method for extracting information from clinical notes using Natural Language Processing methods and the Paragraph Vectors algorithm. Machine Learning algorithms are also used to classify patients with liver, breast, and lung cancer. A comparison and assessment procedure of selected ML models with different parameters was also carried out in order to determine the best one. Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP) are the ML algorithms chosen.

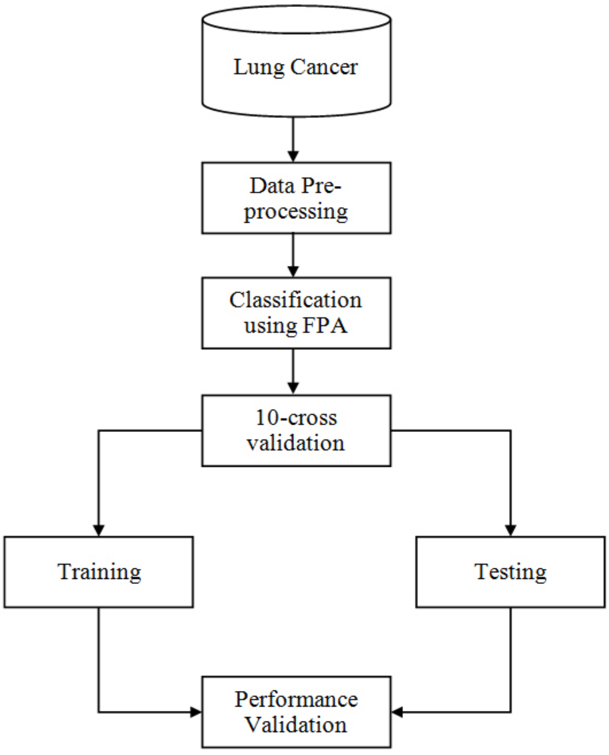
By training Decision Trees (DT) and Random Forests, Venkataraman et al. 2020 defined relevant baseline classification performances (RF). We also looked at whether converting the data with MetaMap Litehad any effect on classification accuracy. The use of LSTM-RNN models is a modular structure that may be helpful in identifying cohorts for oncology studies. Human and veterinary health records will continue to be digitised, especially unstructured narratives. For these two realms to learn from and teach one another, our solution is a step forward. Concept specific identifiers (CUIs) were used by Alawad et al. 2020 as another source of information for the models. With a convolutional neural network (CNN) and a completely linked MLP. By concatenating the high-level document embeddings from text and CUI inputs and then adding a classifier, the high-level document embeddings from text and CUI inputs are merged.

From pathology files, the model is used to remove cancer histology. Kahla et al. (2020) avoided lung cancer by measuring the likelihood of developing the disorder based on the investigated personal knowledge, increasing the main factors of anatomy, and then giving some recommendations to suspect subjects. Our DeepLCP method is built on a mixture of natural language processing (NLP) and CNN. During the validation process of CNN, the DeepLCP experimental findings show a high precision, low error, and loss data rate. Patra (2020) looked at a variety of machine learning classifiers to characterise available lung cancer data in the UCI repository as benign or malignant. The input data is pre-processed and translated to binary format, after which it is classified as cancerous or non-cancerous.

Proposed Method: Many meta-heuristic achieve good accuracy in a variety of research studies and projects related to classification problems in recent years, particularly in binary and multi-classification cases. Furthermore, due to their high performance measure, the FPA method is explored in many other solvable classification problems. As a result, the FPA model is increasingly being used in various meta-heuristic domains, most recently in text classifications (Figure 1). Multi-label classification, on the other hand, is a little more complicated, as it involves label correlation. Multi-label classification is concerned with how to associate an instance with a subset of labels. And dealing with this classification problem is considered can be difficult, particularly when using the traditional simple approach in binary and multi-class cases.

Word2Vec Pre-processing: One of the primary goals of NLP is to extract meaning from a word or text. Using a neural-network approach, methods for embedding words have recently been proposed in case of low-dimensional space. As a result, word embedding is a critical component in a classification problem. This allows us to represent text-words as vectors, which can then be fed into our neural networks. The state-of-the-art word embedding technique is dense vector representation. The words in the text (in this case, toxic text comments) are represented by thick vectors, with each vector representing the projection or mapping into a vector space. Each word position in the vector space is determined by the text within the neighborhood words. A bag-of-words paradigm was historically used, in which vector form of a each word from a whole language is represented by a large sparse vector.

Figure 1: Multi-class Classification



The Google research team was the first to implement Word2vec, with the goal of aggregating similar models to generate word embedding. The Word2vec algorithm they suggested generates Embedding vectors from terms in a text corpus that is more effective than previous methods such as the Latent Semantic Analysis approach. The system computes statistical and frequency of terms in the text corpus, and maps these count-statistics to dense vectors for each given word contained in the corpus in Count-base. Word2vec related models are two-layered, shallow neural networks that are used to construct a linguistic representation of the contexts of the words. It includes a large corpus of text as input and produces a vector-space with hundreds of dimensions, with a vector assigned to each word in the corpus. Furthermore, in the vector space, words with similar meanings in the corpus are clustered together.

Word2vec can also be used as a statistical algorithm, predicting and learning a word embedding from raw text, which is computationally powerful and suitable for mobile and wireless devices (Sumathi A, Saravanan V., 2015). The Skip-Gram model and the Continuous Bag-of-Words model are two architectures for embedding vectors representation in the Word2vec algorithm. In a continuous bag-of-words architecture, the model predicts the target word from source meaning words. The order of the meaning of a word has no bearing on the word prediction in this architecture. Unlike continuous Skipgram architecture, which predicts source context words from a given target word, this model predicts source context words from a given targed word. The remote source context-words are given more weight in this architecture than the close source context-words.

FPA Classification: For the prediction of classes from the given input data to a specific class mark, classification is utilised is of a supervised learning technique. A series of iterative formulae is derived to apply the FPA algorithm. Pollinators such as insects collect improved flower pollen gametes over longer distances in the global pollination step. As a result, the statistical equivalent of improved flower constancy is

$$y_i^{t+1} = y_i^t + \gamma L(\lambda)(y_i^t - y_*)$$

Where,

y_i^{t+1} - solution vector

y_i^t - solution vector at t^{th} iteration,

y_* - current solution,

γ - scaling factor.

$L(\lambda)$ – pollination strength.

The insects or the vectors move for a longer distance, Levy flight distribution can be used with different steps and it reduces the properties. i.e. L> 0. Hence the local pollination condition is represented as below:

$$y_i^{t+1} = y_i^t + \varepsilon(y_i^t - y_k^t)$$

Where

y_j^t - enhanced flower pollen

y_k^t - same plant pollen.

Both the pollens mimics the flow constancy in its neighbourhood and both pollens occurs from similar species. This condition is called as local random walk if uniform distribution ε is in the range between [0, 1]. Pollination from nearby enhanced flowers is more likely to occur in an enhanced flower than pollination from further distant enhanced flowers. To reproduce this, a transfer probability combined with a proximity probability p is used to switch between global and local pollination. According to a provisional parametric, p’=0.8 could be more appropriate for the majority of applications.

RESULTS AND DISCUSSION

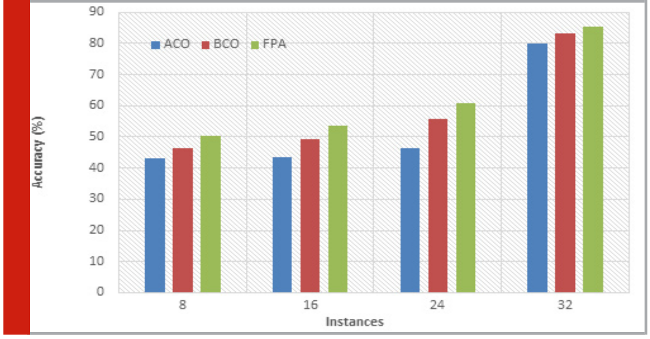
The datasets are collected form UCI repository, where the dataset is found to be a multivariate one with 32 instances and 56 attributes. The performance of the FPA is analysed in terms of carious classification metrics that includes accuracy, sensitivity, specificity, f-measure and mean average percentage error.

The experiments are conducted with the support of Tensor Flow and Keras in Google Colab cloud Service with faster GPU. The other technical specification include 16 GB of available RAM and 20 GB of available free space with 11 generation i7 processor. The simulation is conducted to validate the model, where the behavior of all the performance measures is compared with existing methods namely: bee colony optimization (BCO) and ant colony optimization (ACO) algorithm. Accuracy is defined as the excepted labels vs. the predicted true labels, where the formulation is given below:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{n} I\left(Z_i = Y_i\right)$$

$$where, I\left(true\right) = 1; I\left(false\right) = 0$$

Figure 2: Accuracy



The precision metric estimates the overall percentage of positive classes in the overall classification tasks.

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{n}\frac{\left|Y_i \cap Z_i\right|}{\left|Y_i\right|}$$

Recall is the ability of the classifier to classify the positive instances that are indeed the positive ones.

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{n}\frac{\left|Y_i \cap Z_i\right|}{\left|Y_i \cup Z_i\right|}$$

The F1-score is a weighted harmonic mean of precision and recall, where greatest value is found near to 1 and worst score at 0.

$$F1 = \frac{1}{N}\sum_{i=1}^{n}\frac{2\left|Y_i \cap Z_i\right|}{\left|Y_i\right| + \left|Z_i\right|}$$

Figure 3: Precision

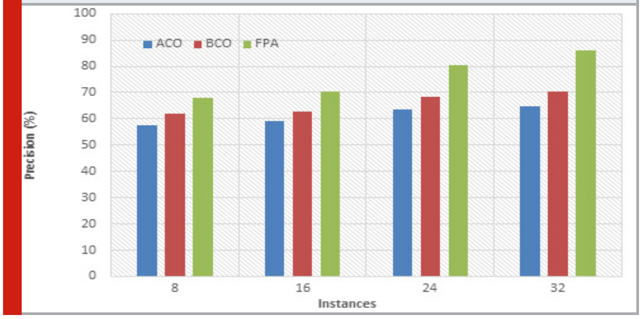


Figure 4: Recall

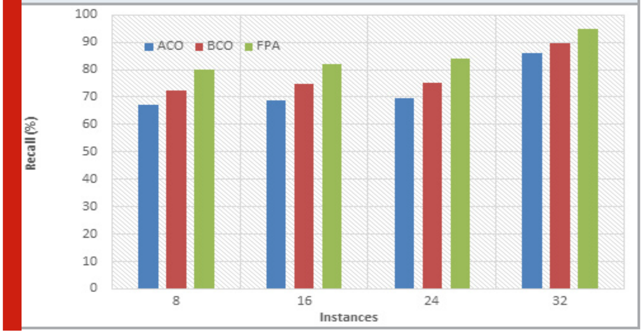


Figure 5: F1-Measure

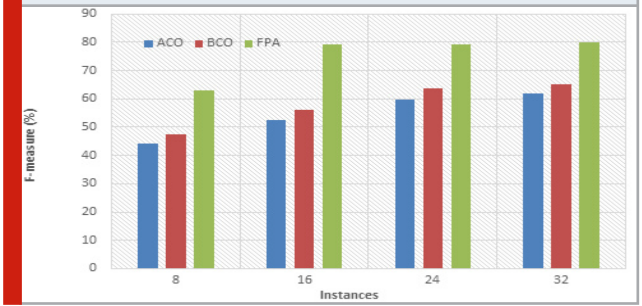
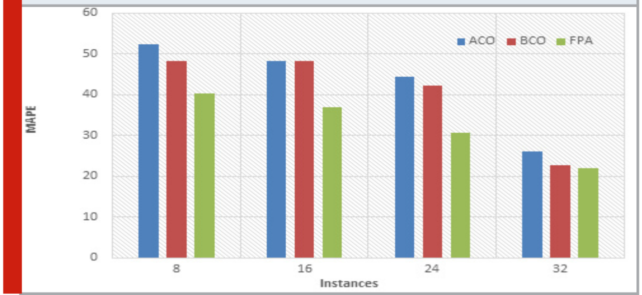


Figure 6: MAPE



The results show that the proposed FPA algorithm (Figure 2- Figure 6) obtains ideal score by assigning each input with each classes based on precision (Figure 3). This helps in penalizing the inaccurate predications of multi-label classes and hence eliminating it. The results of F1-score (Figure 5) outperforms other two models with more than 6.3% with pre-trained word2vec embedding. The results of MAPE (Figure 6) show that the labels that are selected is highly relevant in finding the lung cancer classes. This further contributes to increased recall rate (Figure 4) with unique multi-label classes that helps in easier prediction of lung cancer. At times, premature convergence leads to lower precision that often affects the accuracy (Figure 2) of finding the lung cancer classes.

CONCLUSION

In this paper, the FPA algorithm with series of preprocessing enables the classification of cancer text documents to classify the classes of lung cancer. The FPA framework enables optimal classification of text documents and resolves the issues of multi-label classification. Here, FPA helps in classifying an instant with one or more classes, where a word2vec word embedding approach acts as an embedded corpus. The results show that the classifier performed well even if the datasets are unbalanced. The values of F-measure are higher than other existing models. The comparative results further show that the proposed FPA is effective with increased accuracy of 91.25% over entire dataset than other classifiers. Thus the model is found effective in classifying the multi-label classes than other existing methods with reduced percentage error. In future, the utilization of deep learning can be amended over word2vec model and enable it to perform better word embedding with pre-trained behavior.

REFERENCES

Abdullah, D.M. and Abdulazeez, A.M., 2021. Machine Learning Applications based on SVM Classification A Review. Qubahan Academic Journal, 1(2), pp.81-90.

Agazzi, G.M., Ravanelli, M., Roca, E., Balzarini, P., Pessina, C., Vermi, W., Berruti, A., Maroldi, R. and Farina, D., 2021. CT texture analysis for prediction of EGFR mutational status and ALK rearrangement in patients with non-small cell lung cancer. La radiologia medica, 126(6), pp.786-794.

Alawad, M., Gao, S., Alamudun, F. T., Wu, X. C., Durbin, E. B., Doherty, J., ... & Tourassi, G., 2020. Multimodal Data Representation with Deep Learning for Extracting Cancer Characteristics from Clinical Text. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).

Alzu'bi, A., Najadat, H., Doulat, W., Al-Shari, O., & Zhou, L., 2021. Predicting the recurrence of breast cancer using machine learning algorithms. Multimedia Tools and Applications, 80(9), 13787-13800.

Bright, R., Rankin, S., Dowdy, K., Blok, S.V. and Palmer, L., 2021. Potential Blood Transfusion Adverse Events Can be Found in Unstructured Text in Electronic Health Records using the'Shakespeare Method'. medRxiv.

Debata, P.P., Mohapatra, P., Mishra, D. and Borah, S., 2021. A Chaotic-Jaya Optimized OSELM Model for Cancer Classification. In Soft Computing Techniques and Applications (pp. 611-621). Springer, Singapore.

Garikapati, P., Balamurugan, K., Latchoumi, T.P. and Malkapuram, R., 2021. A Cluster-Profile Comparative Study on Machining AlSi 7/63% of SiC Hybrid Composite Using Agglomerative Hierarchical Clustering and K-Means. Silicon, 13, pp.961-972.

Gowrishankar, J., Narmadha, T., Ramkumar, M. and Yuvaraj, N., 2020. Convolutional Neural Network Classification On 2d Craniofacial Images. International Journal of Grid and Distributed Computing, 13(1), pp.1026-1032.

Kahla, M. B., Kanzari, D., & Maalel, A., 2020. DeepLCP: Towards a DeepLearning Approach to Prevent Lung Cancer. In Digital Health in Focus of Predictive, Preventive and Personalised Medicine (pp. 17-24). Springer, Cham.

Kerr, K.M., Bibeau, F., Thunnissen, E., Botling, J., Ryška, A., Wolf, J., Öhrling, K., Burdon, P., Malapelle, U. and Büttner, R., 2021. The evolving landscape of biomarker testing for non-small cell lung cancer in Europe. Lung Cancer.

Lai-Kwon, J., Heynemann, S., Flore, J., Dhillon, H., Duffy, M., Burke, J., Briggs, L., Leigh, L., Mileshkin, L., Solomon, B. and Ball, D., 2021. Living with and beyond metastatic non-small cell lung cancer: the survivorship experience for people treated with immunotherapy or targeted therapy. Journal of Cancer Survivorship, 15(3), pp.392-397.

Liu, R., Rizzo, S., Whipple, S., Pal, N., Pineda, A.L., Lu, M., Arnieri, B., Lu, Y., Capra, W., Copping, R. and Zou, J., 2021. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature, 592(7855), pp.629-633.

Malika, S. and Jaina, S., Semantic Ontology-Based Approach to Enhance Text Classification. In CEUR Workshop Proceedings (Vol. 2786, pp. 85-98).

Masquelin, A. H., Cheney, N., Kinsey, C. M., & Bates, J. H. 2021. Wavelet decomposition facilitates training on small datasets for medical image classification by deep learning. Histochemistry and Cell Biology, 155(2), 309-317.

Masud, M., Sikder, N., Nahid, A.A., Bairagi, A.K. and AlZain, M.A., 2021. A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. Sensors, 21(3), p.748.

Ma, X., Wang, S., Zhang, Y., Wei, H., & Yu, J., 2021. Efficacy and safety of immune checkpoint inhibitors (icis) in extensive-stage small cell lung cancer (sclc). Journal of Cancer Research and Clinical Oncology, 147(2), 593-606.

Metovic, J., Barella, M., Bianchi, F., Hofman, P., Hofman, V., Remmelink, M., Kern, I., Carvalho, L., Pattini, L., Sonzogni, A. and Veronesi, G., 2021. Morphologic and molecular classification of lung neuroendocrine neoplasms. Virchows Archiv, pp.1-15.

Miller, H.A., Emam, R., Lynch, C.M., Bockhorst, S. and Frieboes, H.B., 2021. Discrepancies in metabolomic biomarker identification from patient-derived lung cancer revealed by combined variation in data pre-treatment and imputation methods. Metabolomics, 17(4), pp.1-13.

Montelongo González, E. E., Reyes Ortiz, J. A., & González Beltrán, B. A., 2020. Machine Learning Models for Cancer Type Classification with Unstructured Data. Computación y Sistemas, 24(2).

Moser, S.S., Bar, J., Kan, I., Ofek, K., Cohen, R., Khandelwal, N., Shalev, V., Chodick, G. and Siegelmann-Danieli, N., 2021. Real world analysis of small cell lung cancer patients: prognostic factors and treatment outcomes. Current Oncology, 28(1), pp.317-331.

Natarajan, Y., Kannan, S. and Mohanty, S.N., 2021. Survey of Various Statistical Numerical and Machine Learning Ontological Models on Infectious Disease Ontology. Data Analytics in Bioinformatics: A Machine Learning Perspective, pp.431-442.

Patra, R., 2020, March. Prediction of Lung Cancer Using Machine Learning Classifier. In International Conference on Computing Science, Communication and Security (pp. 132-142). Springer, Singapore.

Prasath, S., Validating Data Integrity in Steganographed Images using Embedded Checksum Technique. International Journal of Computer Applications, 975, p.8887.

Saravanan V, Mohan Raj V, 2016. Maximizing QoS by cooperative vertical and horizontal handoff for tightly coupled WiMAX/WLAN overlay networks, The Journal of Networks, Software Tools and Applications, Springer, 19(3), pp. 1619-1633.

Saravanan V, Mohan Raj V., 2016. A Seamless Mobile Learning and Tension Free Lifestyle by QoS Oriented Mobile Handoff, Asian Journal of Research in Social Sciences and Humanities, Asian Research Consortium, 6(7), pp. 374-389.

Senthilkumar, P., 2021. Analysis On Industrial Internet Of Things Using Deep Neural Multi-Layer Perceptron Based Model-Based Engineering. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), pp.5550-5555.

Shaikh, S., Daudpota, S.M., Imran, A.S. and Kastrati, Z., 2021. Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models. Applied Sciences, 11(2), p.869.

Sumathi A, Saravanan V., 2015. Bandwidth based vertical handoff for tightly coupled WiMAX/WLAN overlay networks, Journal of Scientific & Industrial Research, vol. 74, pp. 560-566.

Triplette, M., Thayer, J. H., Kross, E. K., Cole, A. M., Wenger, D., Farjah, F., ... & Crothers, K. 2021. The impact of smoking and screening results on adherence to follow-up in an academic multisite lung cancer screening program. Annals of the American Thoracic Society, 18(3), 545-547.

Venkataraman, G. R., Pineda, A. L., Bear Don't Walk IV, O. J., Zehnder, A. M., Ayyar, S., Page, R. L., ... & Rivas, M. A., 2020. FasTag: Automatic text classification of unstructured medical narratives. PloS one, 15(6), e0234647.

Yuvaraj, N., Srihari, K., Chandragandhi, S., Raja, R.A., Dhiman, G. and Kaur, A., 2021. Analysis of protein-ligand interactions of SARS-Cov-2 against selective drug using deep neural networks. Big Data Mining and Analytics, 4(2), pp.76-83.

Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G., & Rajan, A. R., 2021. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. Computers & Electrical Engineering, 92, 107186.

Yuvaraj, N., Srihari, K., Dhiman, G., Somasundaram, K., Sharma, A., Rajeskannan, S., Soni, M., Gaba, G.S., AlZain, M.A. and Masud, M., 2021. Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking. Mathematical Problems in Engineering, 2021.