

International Journal Research Publication Analysis

Page: 01-12

CLUSTERING OF CANCER TEXT DOCUMENTS IN MEDICAL FIELD USING MACHINE LEARNING HEURISTICS

***C. Premila Rosy**

Assistant Professor and Head, PG and Research Department of Computer Science, Idhaya College for Women (Autonomous), Affiliated to Bharathidasan University, Kumbakonam.

Article Received: 06 August 2026

***Corresponding Author: C. Premila Rosy**

Article Revised: 26 August 2026

Assistant Professor and Head, PG and Research Department of Computer Science, Idhaya College for Women (Autonomous), Affiliated to Bharathidasan University, Kumbakonam.

Published on: 16 September 2026

DOI: <https://doi-doi.org/101555/ijrpa.6429>

ABSTRACT

The data clustering over medical text documents plays a major role in extracting relevant information from the documents. However, most of the methods fails in finding the accurate solution on finding the relevant cancer type due to the presence of redundant data items. It is hence necessary to develop a clustering framework that strictly eliminates the redundant data items. In this paper, we present a clustering framework that tends to accurately cluster the cancer text documents to predict what type of cancer is present in a patient. A large database is tested and clustering using the machine learning model. The clustering framework consists of pre-processing the text documents, feature extraction, feature selection and clustering. The clustering using multi-support vector machine enables optimal clustering of text documents. The cancer datasets is used to validate the models over various medline cancer documents dataset. The experimental validation shows improved clustering of documents using the proposed models than other methods.

KEYWORDS: Machine learning, soft computing paradigm, cancer text documents, redundancy reduction.

1. INTRODUCTION

The clustering strategy is generally used to organize a limited collection of clusters, which mostly consist of predefined numbers of clusters, which reflect a dataset based on correlations between its objects [1]. As isolation of overlapping clusters is performed, clustering strategies are of low efficiency. In the partitioning clustering process, meta-

heuristic optimisation algorithms are used. The partial clustering partitions a dataset into a subset of classes dependent on a specific fitness function [17]. The exercise mechanism directly affects the type of group forming. The partitioning method becomes an optimization problem until an adequate fitness function is selected [13]-[17].

Clustering can also be officially viewed from an optimization perspective as a sort of hard-optimization problem (NP) [11]. The need for functional algorithms of optimization has been promoted, which not only provides ad hoc learning for some groups of problems, but also provides the usability of methods of general optimization [12]. Unusually, meta-heuristical and evolutionary optimizing approaches are generally found useful in resolving NP-hard issues, which allows the clustering of problems in reasonable time to produce almost optimal solutions (optimal clusters). According to this assumption, several meta-heuristic algorithms have been implemented in the literature to solve problem clusters, in particular text clusters. These algorithms of optimisation are used for optimal purposes (i.e. fitness function) for the control of search improvement [2]-[7].

In this paper, we present a clustering framework that tends to accurately cluster the cancer text documents to predict what type of cancer is present in a patient. A large database is tested and clustering using the machine learning model. The clustering framework consists of pre-processing the text documents, feature extraction, feature selection and clustering. The clustering using multi-support vector machine enables optimal clustering of text documents. The cancer datasets is used to validate the models over various medline cancer documents dataset.

2. Background

Clusters are a mathematical tool which tries to identify structures or certain patterns in a dataset where there is a certain resemblance to the object inside each cluster. Clustering is a collection of data structures in the same group identical to each other and distinct from objects in the other group. Contrary to the classification, class labels are in classification unknown and groups must be determined by the clustering algorithm [20]-[24]. Clustering is often referred to as unmonitored learning because the clustering does not rely on established classes. The clusters are usually focused on the rules of maximization of similarities between objects of the same classes and minimisation of similarity between objects of different classes (interclass proximity).

Massive amounts of data are collected in the modern era is supposed to double the amount of data received. Gaining data volume information is one of the preferred data mining

attributes.[18]-[19]. There is usually a big difference between the data stored and the information that can be interpreted by the data. This gaps cannot be bridged wherever data extraction enters into the picture immediately. Preliminary information is derived from data in exploratory data studies, but data mining may be helpful by delivering more comprehensive data knowledge. Rapid advancement of IT, engineering and methodology introduces new requirements for the IT paradigm. For some time now, manual data processing has been around, but it causes a significant data analysis bottleneck [25]-[29].

Complex natural world definitions also have a vast range of characteristics. The result of the slowly filled space cannot exponentially increase the amount of points accessible by dimensionality, and the discrimination between clusters is often very weak in full-dimensional space [8]-[10].

Not all dimensions are usually linked in high dimension spaces: data is combined to a specific cluster around those dimensions. Some clusters in the data decrease the frequency of insignificant features. For a clustering process, clusters that are embedded in sub-spaces that are potentially made of various dimensional mixtures in different data locations are determined.

The large sizes and dimensional databases related to the application of data mining need very scalable clustering algorithms. Sampling and parallelization can be used to enhance scalability. Clustering algorithms at a distance prefer to find clusters of the same size and density. Clustering algorithms to detect groups of arbitrary form, density, data coverage, size are essential. This would help to develop a better understanding of the various associations between functions, thereby facilitating additional phases of KDD, for example, decision making.

3. Proposed Method

We carry out the following extraction and clustering model in this section, where the entire data set process is classified in Figure 1.

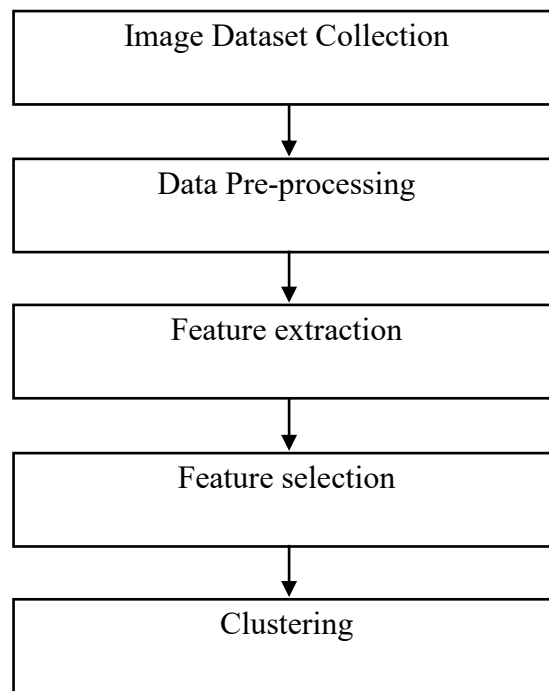


Figure 1: Proposed Classification Model.

a. Pre-Processing

The main aim of clustering is to generate classes according to the intrinsic contents of the objects. If clusters are formed, text requires pre-processing of model steps

Tokenization

Tokenization is the act of splitting words into pieces (words), which are called tokens, possibly lacking characters, such as punctuation concurrently. These tokens are typically linked to terms/words, but distinguishing between type/token is critical.

Stop Words Removal

Standard vocabulary and common phrases and other popular sentences in the language, which are commonly used and useful short words. These words should be omitted from the document (text) provided because they are extremely redundant in general and hence reduce the effectiveness of the clustering techniques.

Stemming

Stemming is the period where a common language is shortened to its source. The stem form is not the same as the morphological root method; it is usually the same stem as describing words, but it is not a true root alone.

Document Representation

The Vector Space Model (VSM) is an important model for defining the contents of documents in an official format. It appeared in the early 1970s. Any paper is designed to

support the estimation of similarities as a term weight vector. Every word in the text transmits a measure of the weighted output, in order to increase the efficiency of the clustering algorithm and to reduce time costs.

b. Feature Extraction

The collection of features is one of the most critical classification system measures. Application filtering is usually used to decrease data size for tens to hundreds of thousands of features that cannot be processed further. The high dimensionality of the function area is a big challenge in web page classification. The best subset of features includes the least number of features that add to precision and reliability of classification.

c. MSVM Classification

The MSVM solves the hyper-plane separation between the groups and it provides the optimal separation between the features in a cancer dataset. If the classification is not linear from the data points, the dataset is divided into multi distinct classes. The data points are transformed by the nonlinear mapping $\phi(x)$ over a high dimensional space that resolves the nonlinear problem between the classes x_i . This helps in separation of points in the solution space using a label y_i . Thus MSVM solves the resultant problem with the feature extracted from the N-point data x_i and it is represented as below:

$$\min_{w, b, \xi_i} 0.5 w^T w + c \sum_{i=1}^N \xi_i \quad (1)$$

$$\text{s.t. } y_i (w^T \phi(x) + b) \geq 1 - \xi_i \quad (2)$$

where,

ξ_i - slack variables

$c \geq 0$ that defines the tradeoff existing between the computational complexity and training error.

A dual Lagrangian function optimizes the MSVM class and it is expressed as below:

$$\min_{\alpha} 0.5 \sum_{i=1}^N \sum_{j=1}^N y_i y_j K(x_i, x_j) \alpha_i \alpha_j - \sum_{j=1}^N \alpha_j \quad 0 \leq \alpha_i \leq C \quad (3)$$

$$\text{s.t. } \sum_{j=1}^N \alpha_j y_j = 0 \quad (4)$$

where

K - kernel vector and it is expressed as below between two different variable x and y :

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \text{ with } \langle \phi(x_i), \phi(x_j) \rangle \quad (5)$$

Hence for a data point x , the predicted class from the text document is given as below:

$$\text{sign}\left(\sum_{j=1}^N \alpha_j y_j K(x, x_j) + b\right) \quad (6)$$

The unwanted data samples are removed at the end of prediction using MSVM that removes the irrelevant support vectors. The prediction limits are hence classified with relevance to clear and most relevant feature set.

4. RESULTS AND DISCUSSIONS

In this section, 85% of the data is used for training the clustering algorithm and 15% of the data is used for testing the algorithm. The performance of the model is tested against medline cancer datasets over different performance metrics that includes accuracy. The proposed model is compared with existing three different classifiers that includes Random Forest, Naïve Bayes, K-nearest neighbor and SVM. The M-SVM classifies the multi-data point to relevant classes and a hyperplane provides the separation between the groups that helps in estimation of relevant clusters.

The accuracy for clustering is expressed as below:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (7)$$

where,

TP is True Positive rate

TN is True Negative rate

FP is False Positive rate

FN is False Negative rate

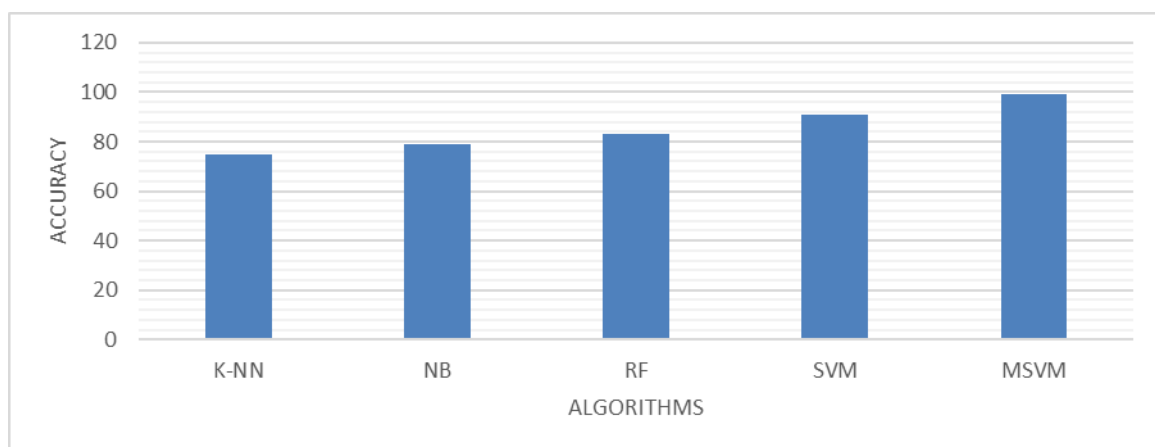


Figure 2: Clustering accuracy on Breast Cancer Wisconsin. (Diagnostic)

Figure 2 displays the clustering accuracy of various algorithms on the Breast Cancer Wisconsin (Diagnostic) dataset. It compares the performance of algorithms such as K-NN, Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Multi-Support Vector Machine (MSVM). The results indicate the effectiveness of each algorithm in accurately clustering the data.

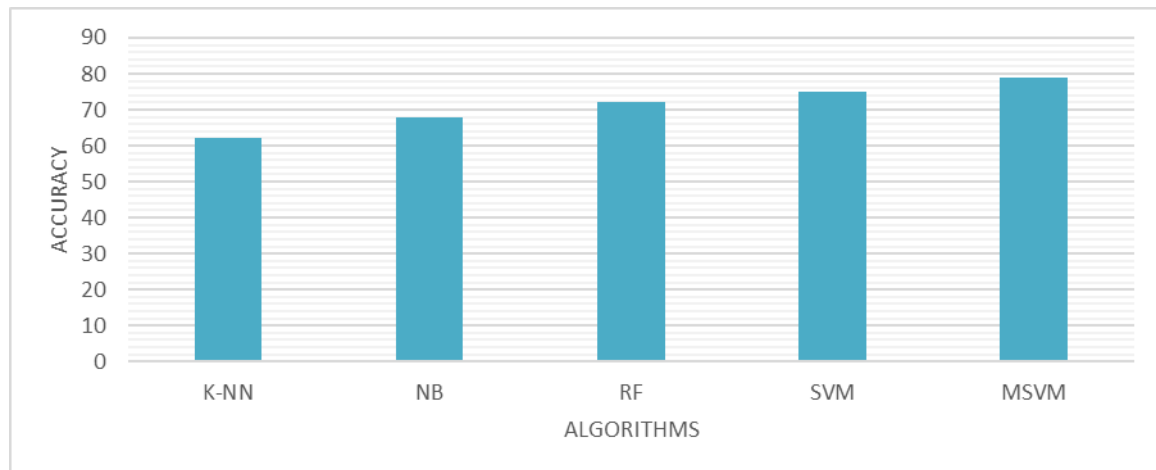


Figure 3: Clustering accuracy on Breast Cancer dataset.

Figure 3 illustrates the clustering accuracy of different algorithms on the Breast Cancer dataset. It provides a comparative analysis of the performance of K-NN, Naïve Bayes (NB), Random Forest (RF), SVM, and MSVM. The figure justifies the effectiveness of these algorithms in accurately clustering the dataset, highlighting their respective accuracy rates.

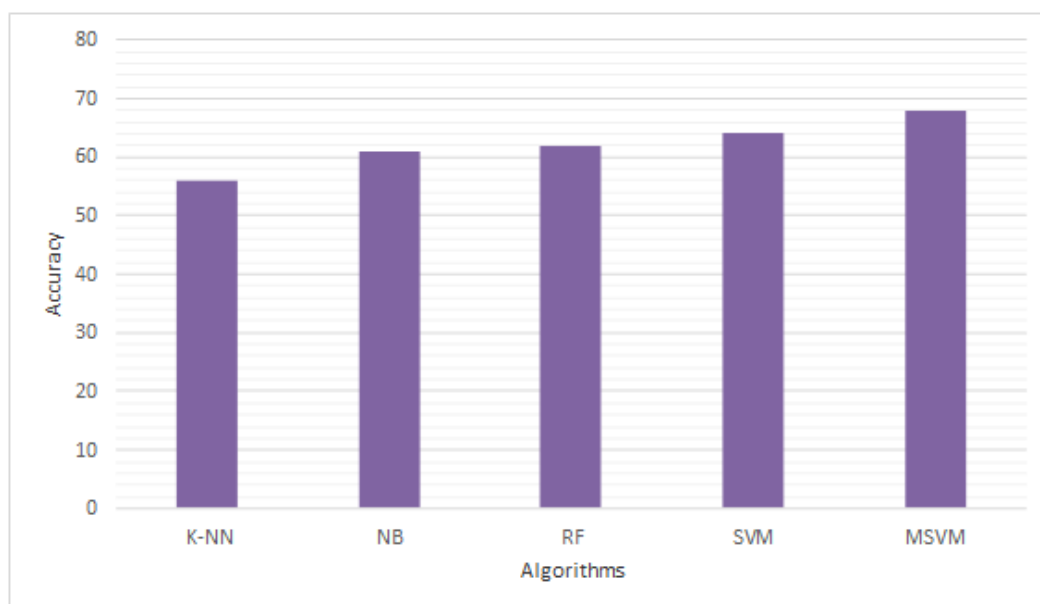


Figure 4: Clustering accuracy on medline cancer dataset.

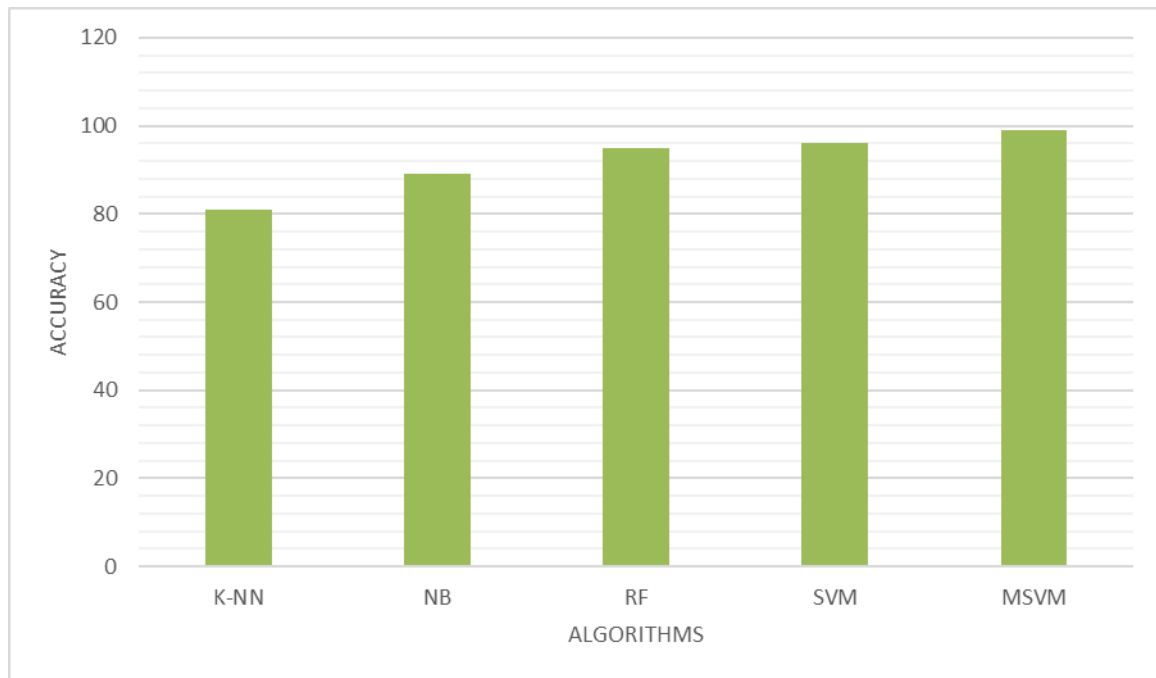


Figure 5: Clustering accuracy on Haberman Breast Cancer Survival Dataset.

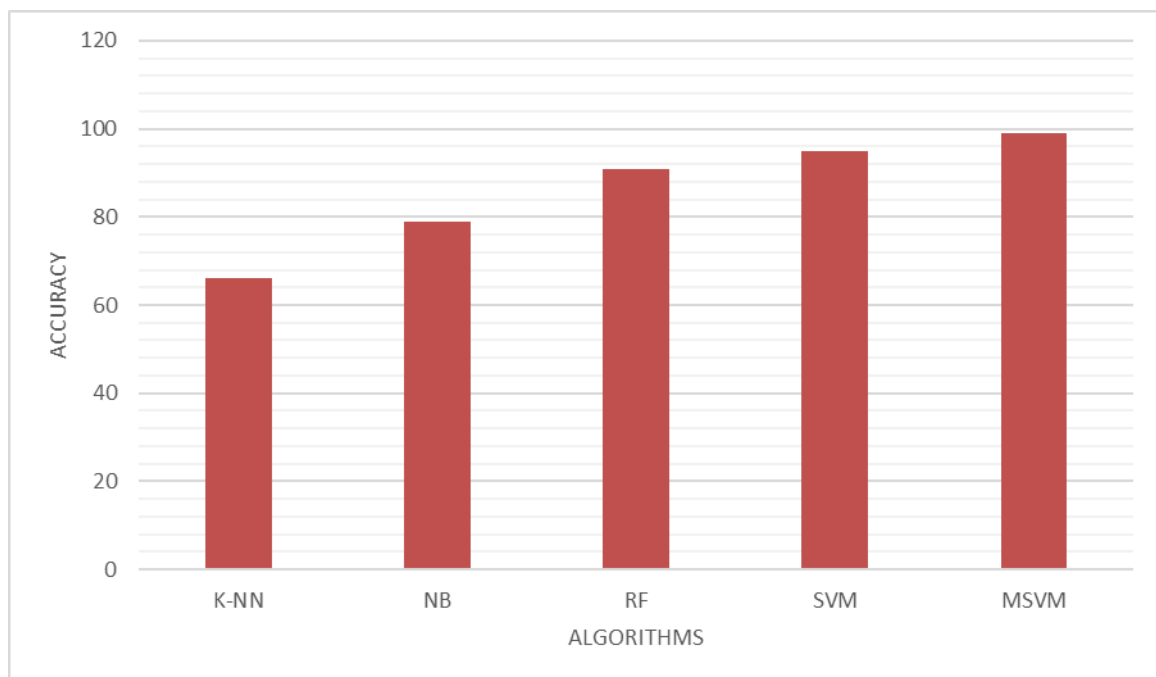


Figure 6: Clustering accuracy on Lung Cancer dataset.

Figure 2 shows the clustering accuracy on Breast Cancer Wisconsin (Diagnostic). Figure 3 shows the clustering accuracy on Breast Cancer dataset. Figure 4 shows the Clustering accuracy on medline cancer dataset. Figure 5 shows the clustering accuracy on Haberman Breast Cancer Survival Dataset. Figure 6 shows the clustering accuracy on Lung Cancer

dataset. From the results, it is seen that the proposed method achieves higher rate of accuracy on all datasets than other methods.

5. CONCLUSIONS

In this paper, we present an M-SVM clustering to cluster accurately the text documents that predicts well the type of cancer in patients. The M-SVM is tested with a large database and with various other machine learning models. The optimal extraction of features at a better rate helps the M-SVM classifiers to classify the type of cancer from the text documents. The validation shows that the proposed method achieves higher rate of accuracy than other existing machine learning classifiers.

REFERENCES

1. Yoon, H. J., Robinson, S., Christian, J. B., Qiu, J. X., & Tourassi, G. D. (2018, March). Filter pruning of convolutional neural networks for text classification: a case study of cancer pathology report comprehension. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (pp. 345-348). IEEE.
2. Nguyen, E., Theodorakopoulos, D., Pathak, S., Geerdink, J., Vijlbrief, O., Van Keulen, M., & Seifert, C. (2020, October). A Hybrid Text Classification and Language Generation Model for Automated Summarization of Dutch Breast Cancer Radiology Reports. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)* (pp. 72-81). IEEE.
3. Jasmir, J., Nurmaini, S., Malik, R. F., & Tutuko, B. (2021). Bigram feature extraction and conditional random fields model to improve text classification clinical trial document. *Telkomnika*, 19(3).
4. Dwivedi, A. K. (2018). Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural Computing and Applications*, 29(12), 1545-1554.
5. Liao, Y., Peng, Y., Liu, D., & Liu, J. (2021, March). Intelligent classification of breast cancer based on deep learning. In *Journal of Physics: Conference Series* (Vol. 1827, No. 1, p. 012171). IOP Publishing.
6. Yoon, H. J., Alawad, M. M., & Tourassi, G. (2018). *Multi-Task Convolutional Neural Networks for Natural Text Classification* (No. Multi-Task Convolutional Neural Networks for Natur; 005841WKSTN00). Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).

7. Baker, S., Ali, I., Silins, I., Pyysalo, S., Guo, Y., Högberg, J., ... & Korhonen, A. (2017). Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics*, 33(24), 3973-3981.
8. Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical text classification using convolutional neural networks. *Stud Health Technol Inform*, 235, 246-50.
9. Kečo, D., Subasi, A., & Kevric, J. (2018). Cloud computing-based parallel genetic algorithm for gene selection in cancer classification. *Neural Computing and Applications*, 30(5), 1601-1610.
10. Gowrishankar, J., Narmadha, T., Ramkumar, M., & Yuvaraj, N. (2020). Convolutional Neural Network Classification On 2d Craniofacial Images. *International Journal of Grid and Distributed Computing*, 13(1), 1026-1032.
11. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2(2), 13.
12. Yuvaraj, N., Raja, R. A., Kousik, N. V., Johri, P., & Diván, M. J. (2020). Analysis on the prediction of central line-associated bloodstream infections (CLABSI) using deep neural network classification. In *Computational Intelligence and Its Applications in Healthcare* (pp. 229-244). Academic Press.
13. Diab, D. M., & El Hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. *Applied Soft Computing*, 54, 183-199.
14. Raja, R. A., & Kousik, N. V. (2021). Privacy Preservation Between Privacy and Utility Using ECC-based PSO Algorithm. In *Intelligent Computing and Applications* (pp. 567-573). Springer, Singapore.
15. Raja, R. A., Karthikeyan, T., & Kousik, N. V. (2020). 11 Improved Privacy Preservation Framework for Cloud-Based Internet of Things. *Internet of Things: Integration and Security Challenges*, 165.
16. Jensen, K., Soguero-Ruiz, C., Mikalsen, K. O., Lindsetmo, R. O., Kouskoumvekaki, I., Girolami, M., ... & Augestad, K. M. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific reports*, 7(1), 1-12.
17. Yuvaraj, N., Srihari, K., Dhiman, G., Somasundaram, K., Sharma, A., Rajeskannan, S., ... & Masud, M. (2021). Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking. *Mathematical Problems in Engineering*, 2021.

18. Jensen, K., Soguero-Ruiz, C., Mikalsen, K. O., Lindsetmo, R. O., Kouskoumvekaki, I., Girolami, M., ... & Augestad, K. M. (2017). Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific reports*, 7(1), 1-12.
19. Jang, B., Kim, M., Harerimana, G., Kang, S. U., & Kim, J. W. (2020). Bi-LSTM model to increase accuracy in text classification: combining Word2vec CNN and attention mechanism. *Applied Sciences*, 10(17), 5841.
20. Ibrahim, M. A., Khan, M. U. G., Mehmood, F., Asim, M. N., & Mahmood, W. (2021). GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification. *Journal of Biomedical Informatics*, 116, 103699.
21. Vilorio, A., Alberto, N., & Pinillos-Patiño, Y. (2021). Classification of Clinical Reports for Supporting Cancer Diagnosis. In *Proceedings of International Conference on Intelligent Computing, Information and Control Systems* (pp. 421-428). Springer, Singapore.
22. Chen, D., Qian, G., & Pan, Q. (2018, December). Breast cancer classification with electronic medical records using hierarchical attention bidirectional networks. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 983-988). IEEE.
23. Ho, C. Y., Syamsudin, M., & Shen, Y. C. (2020, November). Cancer Literature Classification Methods Performance. In *2020 International Conference on Decision Aid Sciences and Application (DASA)* (pp. 801-805). IEEE.
24. Gao, L., Ye, M., & Wu, C. (2017). Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony. *Molecules*, 22(12), 2086.
25. Yao, L., Mao, C., & Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC medical informatics and decision making*, 19(3), 71.
26. Kannan, S., Dhiman, G., Natarajan, Y., Sharma, A., Mohanty, S. N., Soni, M., ... & Gheisari, M. (2021). Ubiquitous Vehicular Ad-Hoc Network Computing Using Deep Neural Network with IoT-Based Bat Agents for Traffic Management. *Electronics*, 10(7), 785.
27. Khadidos, A., Khadidos, A. O., Kannan, S., Natarajan, Y., Mohanty, S. N., & Tsaramirsis, G. (2020). Analysis of COVID-19 Infections on a CT Image Using DeepSense Model. *Frontiers in Public Health*, 8.

28. Kiruthiga, G., Devi, G. U., Raja, R. A., &Kousik, N. V. Analysis of Hybrid Deep Neural Networks with Mobile Agents for Traffic Management in Vehicular Adhoc Networks. Traffic, 16, 18.
29. Karthick, S., Rajakumari, P. A., & Raja, R. A. (2021). Ensemble Similarity Clustering Frame work for Categorical Dataset Clustering Using Swarm Intelligence. In Intelligent Computing and Applications (pp. 549-557). Springer, Singapore.