iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62465

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Medical Symptom Text Classification through NLP in Healthcare

Roji[1], Dr. Deepak Kumar[2]
[1]*Student,* [2]*Assistant professor, CSE, State Institute of Engineering & Technology, Nilokheri, Haryana*

Abstract: ***Medical symptom text classification through Natural Language Processing (NLP) is a rapidly evolving field that aims to leverage computational techniques to analyse and interpret vast amounts of textual data generated in healthcare settings. This paper provides a comprehensive survey of current methodologies, applications, challenges, and future directions in this domain. We begin by discussing the importance of symptom classification for improving patient outcomes, supporting clinical decision-making, and enhancing disease surveillance. We then review traditional machine learning approaches and advanced deep learning models, highlighting their respective strengths and limitations. Key pre-processing techniques crucial for handling medical jargon and ensuring data privacy are also examined. The paper further explores real-world applications, including clinical decision support systems, disease outbreak detection, and patient monitoring. Despite significant advancements, challenges such as data quality, model interpretability, and regulatory compliance remain. Finally, we identify emerging trends and potential future developments that could drive further innovation in NLP for healthcare. This survey aims to provide a valuable resource for researchers and practitioners seeking to understand and contribute to the field of medical symptom text classification.***

I. INTRODUCTION

1) Background on the Significance of Medical Symptom Text Classification

Medical symptom text classification is vital for accurately interpreting and categorizing symptoms described in clinical notes, patient records, and other textual data. This classification enhances clinical decision-making, improves patient care, and streamlines medical record management by automating the extraction and analysis of critical health information.

2) The Role of NLP in Transforming Healthcare Data Analysis

Natural Language Processing (NLP) transforms healthcare data analysis by enabling the automated extraction and interpretation of unstructured text data. NLP techniques support clinical decision-making, enhance patient care, streamline administrative processes, and facilitate medical research. By processing large volumes of textual data efficiently, NLP helps uncover insights that improve patient outcomes and operational efficiencies in healthcare.

3) Objectives and Contributions of the Survey

The survey aims to provide a comprehensive overview of current techniques and advancements in medical symptom text classification using NLP. It highlights the challenges, methodologies, and applications in the field, offering insights into future research directions. The contributions include a detailed analysis of existing approaches, identification of gaps in current research, and recommendations for enhancing NLP applications in healthcare.

II. LITERATURE REVIEW

Author Name	Study Title	Methodology	Key Findings
Y. Wang et al. [2023]	Medical text classification based on the discriminative pre-	Discriminative pre-training, NLP techniques, Machine Learning algorithms	Developed a discriminative pre-training model for medical text classification, achieving high accuracy in categorizing medical texts
B. Zhou et	Medical Text	Improved BI-LSTM,	Proposed a deep learning model combining BI-LSTM and

©IJRASET: All Rights are Reserved

Author Name	Study Title	Methodology	Key Findings
al. [2021]	Classification System Based on Deep Learning	Attention Mechanism	attention mechanism for accurate medical text classification
L. Yao et al. [2019]	Clinical text classification with rule-based features and ...	Rule-based features, Machine Learning algorithms	Explored rule-based features and ML algorithms for clinical text classification, achieving promising results
S.K. Prabhakar et al. [2021]	Medical Text Classification Using Hybrid Deep Learning	Hybrid deep learning approach	Developed a hybrid deep learning model for medical text classification, achieving improved performance
T.A. Koleck et al. [2019]	Natural language processing of symptoms documented in ...	NLP techniques, Symptom characterization	Analysed symptoms documented in EHR narratives using NLP, highlighting their role in disease characterization
Q. Zhang et al. [2022]	Research on Medical Text Classification Based ...	NLP techniques, Medical text classification algorithms	Investigated various NLP-based methods for medical text classification, presenting promising results

III. NLP TECHNIQUE IN HEALTHCARE

1) General Introduction to NLP:

Natural Language Processing (NLP) is a field of artificial intelligence focused on enabling computers to understand, interpret, and generate human language. In healthcare, NLP techniques are applied to analyse unstructured text data from various sources like clinical notes, patient records, and medical literature.

2) Specific NLP Methods Used in Healthcare:

a) Text pre-processing:

- Tokenization: Breaking text into individual words or tokens.
- Stemming: Reducing words to their root form by removing suffixes.
- Lemmatization: Similar to stemming but produces valid words

These pre-processing steps standardize and clean text data for further analysis.

b) Named Entity Recognition (NER):

- Identifies and classifies named entities such as diseases, symptoms, medications, and anatomical terms in text.
- NER helps extract structured information from unstructured medical text, facilitating tasks like clinical coding and information retrieval.

c) Text Classification Algorithms:

- Supervised Learning: Trained on labelled data to classify text into predefined categories (e.g., disease diagnosis, medication adherence).
- Unsupervised Learning: Discovers patterns and structures in text data without predefined labels, useful for clustering similar documents or discovering topics.
- Semi-supervised Learning: Combines labelled and unlabelled data to improve classification accuracy, especially when labelled data is limited.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

d) Word Embedding's and Representation Learning:

- Word2Vec, GloVe: Techniques to represent words as dense vectors in a continuous vector space.
- Bidirectional Encoder Representations from Transformers(BERT): Pre-trained deep learning model capturing bidirectional context from text, widely used for various NLP tasks including text classification and named entity recognition.

These approaches capture semantic meanings and relationships between words, enhancing the understanding of medical text by algorithms.

IV. MEDICAL SYMPTOM TEXT CLASSIFICATION METHODS

A. Rule –based Methods:

1) Expert System and Heuristic Rule: Utilize predefined rules based on domain knowledge to classify medical symptoms. These systems rely on logical statements and decision rules crafted by experts in the field. While they are interpretable and straightforward, they may lack adaptability to complex and diverse symptom data.

B. Machine Learning Approaches:

1) Traditional Machine Learning Classifiers: Algorithms like Support Vrector Machines (SVM), Naive Bayes, and Decision Trees are trained on labeled symptom data to classify new instances. These models learn patterns from the data and generalize to unseen examples. However, they require extensive feature engineering and may struggle with complex relationships in medical text data.

2) Feature Engineering Specific to Medical Texts: Involves crafting features tailored to medical text data, such as symptoms, diagnoses, or medical concepts. Feature engineering enhances the performance of machine learning models by capturing relevant information and improving classification accuracy.

C. Deep Learning Approaches:

1) Neural Network: Deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory networks (LSTMs) are capable of learning intricate patterns and relationships from raw text data. They excel at capturing sequential dependencies and semantic information.

2) Attention Mechanism and Transformers: Attention mechanisms, as seen in models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-Trained Transformer), focus on relevant parts of the input text, enabling more accurate classification by attending to important symptom features.

3) Transfer Learning in Medical Text Classification: Transfer learning leverages pre-trained models on large text corpora and fine-tunes them on medical symptom text data. This approach enables effective utilization of general knowledge from large datasets to improve classification performance in medical contexts.

These methods provide diverse approaches to medical symptom text classification, each with its strengths and limitations. Their selection depends on factors such as dataset size, complexity, interpretability, and computational resources available for the task at hand.

V. DATASETS AND EVALUATION METRICS

A. Overview of Commonly Used Datasets:

1) Publicly Available Datasets: Common datasets for medical symptom text classification include MIMIC (Medical Information Mart for Intensive Care), i2b2 (Informatics for Integrating Biology and the Bedside), and OHDSI (Observational Health Data Sciences and Informatics). These datasets contain annotated medical records, clinical notes, and symptom descriptions, facilitating research and development in healthcare NLP tasks.

2) Challenges in Datasets Acquisition and Annotation: Acquiring and annotating medical datasets pose significant challenges due to issues like data privacy, variability in clinical documentation practices, and the need for domain expertise. Annotation requires trained annotators and can be time-consuming and expensive, limiting the availability of labelled data.

B. Evaluation Metrics:

1) Accuracy: Measures the overall correctness of the classification model by calculating the ratio of correctly predicted instances to the total number of instances.

2) *Precision:* Indicates the proportion of correctly predicted positive cases out of all instances predicted as positive. It measures the model's ability to avoid false positives.
3) *Recall:* Measures the proportion of correctly predicted positive cases out of all actual positive instances. It assesses the model's ability to identify all relevant instances.
4) *F1-score:* Harmonic mean of precision and recall, providing a balance between the two metrics. It accounts for both false positives and false negatives, making it suitable for imbalanced datasets.

C. *Specific Metrics Relevant to Healthcare Application:*
1) *Sensitivity (True Positive Rate):* Measures the proportion of actual positive cases correctly identified by the model. It is crucial for detecting diseases or medical conditions.
2) *Specificity (True Negative Rate):* Measures the proportion of actual negative cases correctly identified by the model. It is essential for ruling out conditions and minimizing false alarms.
3) *Area Under the Receiver Operating Characteristics Curve (AUC-ROC):* Evaluates the performance of the classification model across various thresholds. It quantifies the model's ability to distinguish between classes and is commonly used in diagnostic applications.

These evaluation metrics provide quantitative measures to assess the performance of medical symptom text classification models, ensuring their effectiveness and reliability in healthcare applications.

VI. APPLICATION AND CASE STUDIES

A. *Real-World Applications of Symptom Text Classification:*
1) *Clinical Decision Support System:* Symptom text classification is integral to clinical decision support systems, aiding healthcare providers in diagnosing diseases, recommending treatments, and predicting patient outcomes based on symptoms documented in medical records and clinical notes.
2) *Patient Self-Reporting Tools:* Symptom text classification enables the development of patient self-reporting tools, allowing individuals to describe their symptoms electronically. These tools facilitate remote monitoring, patient engagement, and early detection of health issues.
3) *Early Diagnosis and Disease Prediction:* By analysing symptom descriptions from patients, symptom text classification supports early diagnosis and prediction of diseases. It helps identify patterns and trends in symptoms, enabling proactive healthcare interventions and disease management.

B. *Case Studies Demonstrating Successful Implementations:*
1) *Example 1:* A clinical decision support system implemented in a hospital setting utilizes symptom text classification to assist physicians in diagnosing rare diseases. The system analyses symptoms reported by patients and provides recommendations based on historical patient data and medical literature, resulting in more accurate diagnoses and improved patient outcomes.
2) *Example 2:* A mobile application allows patients to self-report their symptoms using natural language. Symptom text classification algorithms process the input text, categorizing symptoms into relevant medical conditions. The application provides personalized health recommendations and alerts patients to seek medical attention if necessary, enhancing patient engagement and self-care.
3) *Example 3:* Researchers develop a predictive model for early detection of heart disease using symptom text classification. By analysing electronic health records and symptom descriptions, the model identifies patterns indicative of cardiac conditions. Early diagnosis enables timely interventions and improves prognosis for patients at risk of heart disease.

These applications and case studies demonstrate the diverse uses and effectiveness of symptom text classification in healthcare, ranging from clinical decision support to patient empowerment and disease prediction. They highlight the value of NLP techniques in leveraging textual data for improved healthcare outcomes.

VII. CHALLENGES

1) *Data Privacy and Ethical Concerns:* NLP models require access to sensitive patient data, raising concerns about privacy and confidentiality. Ensuring compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act) is essential to protect patient privacy and maintain trust in healthcare systems.

2) *Handling Ambiguous and Rare Symptoms:* NLP models may struggle to interpret ambiguous or rare symptoms accurately, leading to misclassifications or errors in diagnosis. Addressing this challenge requires the development of robust algorithms capable of understanding context and capturing nuances in symptom descriptions.
3) *Integrating Multimodal Data:* Healthcare data often include multiple modalities, such as text, images, and lab results. Integrating these diverse data sources poses technical challenges for NLP models. Advancements in multimodal learning techniques are needed to effectively leverage these data types for comprehensive symptom classification.

VIII. FUTURE RESEARCH DIRECTIONS

1) *Advances in NLP Models and their potential Impact:* Future research should focus on developing more advanced NLP models tailored to medical symptom classification tasks. This includes exploring novel architectures, improving model performance on specific healthcare domains, and addressing scalability issues to handle large-scale datasets.
2) *Improved Interpretability and Transparency in models:* Enhancing the interpretability and transparency of NLP models is crucial for gaining trust from healthcare professionals and patients. Research efforts should prioritize developing interpretable models that provide insights into decision-making processes and enable clinicians to understand model predictions.
3) *Collaboration Between Healthcare professionals and Data Scientists:* Collaboration between healthcare professionals and data scientists is essential for bridging the gap between NLP research and clinical practice. This collaboration fosters the development of clinically relevant solutions, ensures alignment with healthcare needs, and promotes responsible deployment of NLP technologies in healthcare settings.

These future research directions aim to overcome current limitations in NLP for medical symptom classification, paving the way for more accurate, interpretable, and ethically sound applications in healthcare.

IX. CONCLUSION

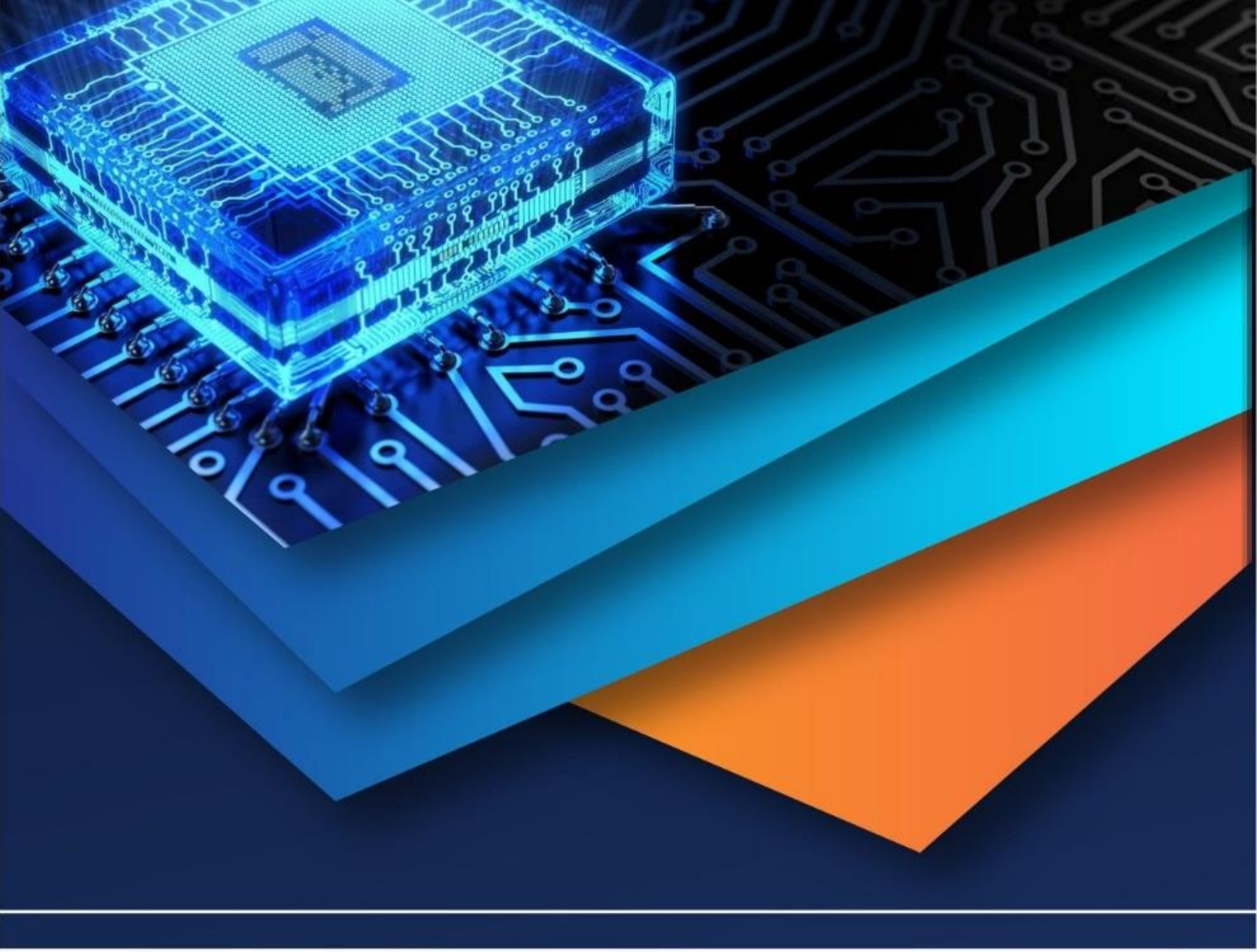
In conclusion, medical symptom text classification using Natural Language Processing (NLP) holds immense potential to revolutionize healthcare by enabling automated analysis and interpretation of unstructured textual data. Through this technology, healthcare providers can enhance clinical decision-making, improve patient care, and facilitate medical research. However, several challenges remain, including data privacy concerns, handling ambiguous symptoms, and integrating multimodal data. Despite these challenges, the future of medical symptom text classification through NLP looks promising. Ongoing research efforts aim to address current limitations by advancing NLP models, improving interpretability, and fostering collaboration between healthcare professionals and data scientists. By leveraging innovative approaches and fostering interdisciplinary collaboration, NLP can continue to drive meaningful advancements in healthcare, ultimately leading to better patient outcomes and more efficient healthcare delivery. In conclusion, medical symptom text classification through NLP represents a transformative technology with the potential to revolutionize healthcare practices and improve patient outcomes in the years to come.

REFERENCES

[1] Medical text Classification system based on deep learning. (2021b, June 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9653548
[2] Yao, L., Mao, C., & Luo, Y. (2019b). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Medical Informatics and Decision Making, 19(S3). https://doi.org/10.1186/s12911-019-0781-4
[3] Prabhakar, S. K., & Won, D. (2021b). Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention. Computational Intelligence and Neuroscience, 2021, 1–16. https://doi.org/10.1155/2021/9425655
[4] Zhang, Q., Qihao, Y., Lv, P., Zhang, M., & Lv, L. (2022). Research on medical text classification based on improved capsule network. Electronics, 11(14), 2229. https://doi.org/10.3390/electronics11142229
[5] Richter-Pechanski P, Geis NA, Kiriakou C. et al. Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. Digital Health 2021.
[6] Altman R. Artificial intelligence (AI) systems for interpreting complex medical datasets. Clin Pharmacol Ther 2017
[7] Névéol A, Dalianis H, Velupillai S. et al. Clinical natural language processing in languages other than English: Opportunities and challenges. J Biomed Semantics 2018
[8] Saad E, Sadiq S, Jamil R. et al. Predicting death risk analysis in fully vaccinated people using novel extreme regression-voting classifier. Digital Health 2022
[9] Mujtaba G, Shuib L, Idris N. et al. Clinical text classification research trends: systematic literature review and open issues. Expert Syst Appl 2019
[10] Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015
[11] Yi K, Beheshti J. A hidden Markov model-based text classification of medical documents. J Inform Sci 2009
[12] Yahia HS, Abdulazeez AM. et al. Medical text classification based on convolutional neural network: a review. Int J Sci Bus 2021
[13] Lavanya P, Sasikala E. Deep learning techniques on text classification using natural language processing (NLP) in social healthcare network: a comprehensive survey. In: 2021 3rd International Conference on Signal Processing and Communication (ICPSC).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[14] JDM W C Kenton and Toutanova LK. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp.4171–4186.
[15] Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, et al.Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLOS ONE. 2018; 13(2):e0192360. https://doi.org/10.1371/journal.pone.0192360.
[16] Luo Y. Recurrent neural networks for classifying relations in clinical notes. J Biomed Inform. 2017; 72:85–95.
[17] Luo Y, Cheng Y, Uzuner Ö, Szolovits P, Starren J. Segment convolutional neural networks (seg-cnns) for classifying relations in clinical notes. J Am Med Inform Assoc. 2017; 25(1):93–8.
[18] Guo, B.; Zhang, C.; Liu, J.; Ma, X. Improving text classification with weighted word embeddings via a multi-channel TextCNN model. Neurocomputing **2019**, 363, 366–374.
[19] Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys. D Nonlinear Phenom. **2020**, 404, 132306.
[20] Jasmir, J.; Nurmaini, S.; Malik, R.F.; Tutuko, B. Bigram feature extraction and conditional random fields model to improve text classification clinical trial document. Telkomnika **2021**,
[21] X. Zhang, R. Henao, Z. Gan, Y. Li, and L. Carin, “Multi-label learning from medical plain text with Convolutional residual models,” Computer Science, vol. 9, 2018



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)