

A Survey on Machine Learning Techniques for Medical Report Analysis

Mr. Shashidhara H V¹, Ramsha Taskeen², Rishitha G D³, Shashank K⁴, Shivaprasad H V⁵

^{1,2,3,4,5} Department of Computer Science and Engineering Malnad College of Engineering
Hassan, Karnataka, India

Email: hvs@mcehassan.ac.in, ramshtsk9972@email.com, gdrishitha4@email.com,
shashankk1201200@email.com shashankk1201200@e

Abstract—It has been observed in recent years that the need for the analysis of medical reports has become very significant due to the generation of a large amount of medical data on a daily basis in hospitals and other diagnostic centers. Manual analysis of medical reports is not only laborious but also expensive and highly error-prone, especially while working with large patient databases. ML/AI is being used in order to analyze medical reports, predict various types of diseases, generate diagnostic information, and assist healthcare professionals during clinical decision-making processes. With the aid of the latest developments in deep learning, NLP, and AI, the efficiency of medical reports analysis has been highly improved. In this survey, an insight into the research conducted in relation to medical report analysis between 2021 and 2026 has been presented. In particular, the utilization of several methods including support vector machine, random forest, convolutional neural networks, recurrent neural network, long short-term memory, and transformers has been thoroughly discussed.

Index Terms—Machine Learning, Medical Report Analysis, Deep Learning, Healthcare Analytics, Radiology Reports, Disease Prediction, NLP, Artificial Intelligence

I. INTRODUCTION

A lot of data is being produced on a daily basis through laboratory tests, radiology tests, patient files, medication prescribed by doctors, and other clinical files. It becomes hard for healthcare professionals to go through all the available information and analyze it since reading through these files consumes much time especially when there are too many patients in the hospital. It may also be difficult to make sense out of this data even when computer programs are utilized in analysis of simple reports.

Machine learning is a type of artificial intelligence technology that is applied to help in the analysis of the medical reports and aid in decision making for the doctors. Various functions can be performed using this technology. One of the things that can be done using machine learning is that hidden patterns in healthcare data can be identified and help physicians diagnose diseases.

Analyzing reports is not an easy task as it involves a series of processes including predicting the diagnosis, classification of reports, preparation of radiology reports, image analysis and extraction of information from texts among other things. There is need for application of NLP methods in this case.

There have been developments such as the emergence of new technologies such as BERT that enhance comprehension of medical reports. Many hospitals are also starting to integrate Artificial Intelligence technologies in order to minimize errors

in their systems and improve the quality of care. Artificial Intelligence has several advantages especially in terms of handling patients' data.

The objective of this paper will be to study machine learning techniques applied in the analysis of medical reports between the years 2021 and 2026. This will include discussion on algorithms, datasets, applications, comparisons, challenges and future research directions in healthcare systems. The topic of discussion will be the analysis of medical reports and machine learning techniques applied in medical report analysis.

II. LITERATURE SURVEY

Medical report analysis has become an active research area due to the rapid growth of healthcare data and advancements in machine learning technologies. Researchers have proposed several ML and DL-based methods for automating clinical report generation, disease prediction, medical image analysis, and healthcare decision support systems.

Sun et al. presented a review on generative AI methods for lesion localization and automatic report generation. The paper discussed multimodal learning frameworks that integrate radiology images with textual clinical information. The study highlighted the effectiveness of attention mechanisms and vision-language models in improving diagnostic report generation. The paper concludes that generative AI systems were shown to achieve better localization of abnormalities and more accurate medical report descriptions compared to conventional deep learning approaches.

Vavekanand et al. presented a comprehensive survey on the application of multimodal machine learning techniques in predictive healthcare analytics. The study focused on integrating multiple healthcare data sources such as medical imaging, clinical text, electronic health records, and time-series physiological signals to improve disease prediction and healthcare decision-making. The survey highlighted important applications in oncology, neurology, and cardiology, where combining diverse healthcare data enhanced diagnostic and prognostic performance. However, the study also identified major challenges such as complex preprocessing, synchronization of heterogeneous data sources, missing data handling, and limited external validation, which affect the practical deployment of the multimodal healthcare systems.

Liu et al. reviewed state-of-the-art deep learning methods for automatic medical report generation and categorized existing

systems into template-based, retrieval-based, and generation-based approaches. Template-based methods use predefined report structures, retrieval-based systems generate reports by retrieving similar historical cases, while generation-based methods use deep learning architectures for producing entirely new reports. The study highlights that the generation-based transformer models significantly improve report fluency and semantic accuracy. The authors also discussed the importance of multimodal learning systems that combine medical imaging data and clinical text information to enhance diagnostic performance. They concluded that the transformer models perform better than the old RNN systems.

Kulsoom et al. reviewed recent advancements in transformer-based machine learning models for radiology report analysis. The study focused on contextual representation learning using BERT and other transformer architectures. BERT helps understand the meaning of sentences and context of words. The authors explained how transformer models capture long-range dependencies in clinical text and improve report classification accuracy. The paper also discussed explainable AI approaches for improving transparency and trustworthiness in healthcare systems. The study concluded that transformer-based NLP systems outperform traditional ML methods in healthcare text analysis tasks.

Hansen et al. reviewed multimodal representation learning methods used in predictive healthcare analytics. The study explained how machine learning models combine multiple healthcare data sources such as patient history, laboratory reports, imaging data, and sensor data for improving disease prediction accuracy. The authors emphasized that multimodal learning systems provide better clinical insights and reduce diagnostic errors compared to single-modal systems.

Pang et al. presented a comprehensive survey on automatic medical imaging report generation using deep learning techniques. The authors discussed various architectures including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), hierarchical attention networks, and transformer-based models. The study explained how CNN models are used for extracting image features from radiology scans while RNN and LSTM models are applied for generating textual diagnostic reports. The paper also compared multiple public datasets such as MIMIC-CXR and IU X-ray datasets and highlighted evaluation metrics including BLEU, ROUGE, and CIDEr scores for report generation systems. The authors concluded that transformer-based architectures provide better contextual understanding and improved report quality compared to traditional methods.

Casey et al. conducted a systematic review on Natural Language Processing (NLP) applications in radiology report analysis. The study focused on extracting structured information from unstructured clinical text documents using machine learning and NLP techniques. Various text mining approaches such as tokenization, stemming, named entity recognition, and feature extraction were discussed. The paper highlighted that NLP-based systems improve radiology workflow efficiency by automatically identifying diseases, abnormalities, and di-

agnostic observations from radiology reports. The authors also emphasized challenges related to domain-specific medical terminology and data standardization.

Shurrab and Duwairi conducted a survey on self-supervised learning methods used in medical imaging analysis. The study explained that self-supervised learning reduces the dependency on large annotated datasets by enabling models to learn feature representations from unlabeled medical data. The paper discussed various contrastive learning and pretraining methods that improve model generalization and robustness in healthcare applications.

Van der Velden et al. analyzed Explainable Artificial Intelligence (XAI) methods in deep learning-based medical image analysis. The study emphasized the importance of explainability in healthcare applications where doctors need to understand the reasoning behind AI predictions. Various explainability methods such as saliency maps, Grad-CAM, SHAP, and LIME were discussed for visualizing model decisions. The authors concluded that XAI techniques improve trust and acceptance of AI systems in clinical environments.

Messina et al. reviewed deep learning and explainability techniques for automatic medical report generation systems. The paper considered the incorporation of XAI approaches into models for generating radiology reports as well as the role of interpretability and explainability of the AI systems under consideration for their successful implementation in healthcare practice. It can be noted that literature produced during the years of 2024 through 2026 pays special attention to using state-of-the-art LLMs, multimodal transformers, and generative AI models for analyzing healthcare-related data. In particular, these models incorporate manipulations with NLP, computer vision, and multimodal learning in order to generate accurate medical reports. Data privacy and security is another crucial aspect of this technology. Moreover, federated learning models that keep patient data decentralized have been used for healthcare analytics purposes in order to protect patients' personal information. As for the current state of machine learning and deep learning in the context of health report analysis, it can be said that many achievements have been reached in this area. However, issues concerning data privacy, lack of appropriate datasets, model interpretability, and high computational complexity need to be studied further.

III. COMPARISON OF LITERATURE SURVEY

The reviewed literature highlights the growing use of machine learning, deep learning, and multimodal learning techniques in medical report analysis. Earlier studies mainly focused on traditional machine learning and NLP methods, while recent research emphasizes transformer architectures, explainable AI, multimodal systems, and generative models. Table II presents a comparative summary of the reviewed studies based on techniques, datasets, applications, advantages, and limitations.

IV. MACHINE LEARNING TECHNIQUES USED IN MEDICAL REPORT ANALYSIS

Machine learning techniques play an important role in automating medical report analysis and disease diagnosis. Different algorithms are used depending on the type of medical data and application requirements.

A. Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm widely used for classification tasks in healthcare systems. SVM is highly effective for disease prediction and clinical text classification because it performs well with high-dimensional datasets. In medical report analysis, SVM models are used for identifying diseases such as cancer, diabetes, and heart disease from patient reports.

The basic principle of SVM is to identify an optimal hyperplane that separates different classes with maximum margin.

SVM models provide high classification accuracy and are commonly applied in healthcare analytics.

B. Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction performance and reduce overfitting. It is widely used for disease prediction and patient risk analysis. This also increases the accuracy of prediction.

Random Forest models can handle large healthcare datasets efficiently and provide robust predictions even with missing data. In medical applications, Random Forest algorithms are commonly used for predicting chronic diseases and analyzing laboratory reports.

C. Convolutional Neural Networks (CNN)

Convolutional Neural Networks are deep learning models mainly used for medical image analysis. CNN models automatically extract important image features and improve disease detection accuracy from X-rays, MRI scans, CT scans, and ultrasound images.

CNN-based architectures are widely adopted in radiology report generation systems where medical images are converted into diagnostic textual reports. Deep CNN models significantly reduce manual feature engineering requirements. Deep CNN architectures have shown strong performance in radiology report generation systems.

D. Recurrent Neural Networks (RNN)

Recurrent Neural Networks are designed for sequential data analysis and are highly effective for processing clinical text reports. RNN models capture contextual relationships between words in medical documents.

Hierarchical RNN architectures are frequently used in automatic medical report generation systems. These models generate meaningful diagnostic reports from radiology images and clinical records.

E. Long Short-Term Memory (LSTM)

LSTM is an advanced form of RNN designed to overcome the vanishing gradient problem. LSTM networks can remember long-term dependencies in sequential medical data.

LSTM models are widely used for:

- Clinical text classification
- Medical report generation
- Disease progression prediction
- Patient monitoring systems

F. Transformer Models

Transformer-based architectures such as BERT and GPT models have recently become highly popular in healthcare NLP applications. Transformers improve contextual understanding of medical text and achieve better performance in clinical report analysis.

Transformer models are widely used for:

- Clinical Named Entity Recognition
- Medical Question Answering
- Radiology Report Generation
- Disease Classification

These models outperform traditional machine learning techniques in many healthcare NLP tasks.

V. DATASETS USED IN MEDICAL REPORT ANALYSIS

Datasets play a crucial role in training and evaluating machine learning models for healthcare applications. Several publicly available medical datasets are widely used by researchers.

TABLE I
COMMON MEDICAL DATASETS

Dataset	Purpose
MIMIC-CXR	Chest X-ray reports and images
IU X-ray	Radiology report generation
CheXpert	Disease classification from X-rays
OpenI	Medical image-text dataset
Heart Disease Dataset	Heart disease prediction
Diabetes Dataset	Diabetes prediction
Breast Cancer Dataset	Cancer classification

VI. APPLICATIONS OF MACHINE LEARNING IN MEDICAL REPORT ANALYSIS

Machine learning algorithms find extensive application across diverse areas in healthcare.

A. Disease Prediction

ML algorithms help predict various diseases such as diabetes, cancer, heart disease, and pneumonia based on data collected and analyzed from laboratory reports and patient information.

B. Radiology Report Generation

Deep Learning algorithms can automatically produce reports for radiology through analysis of X-rays, MRI, ultrasound, and CT scan images.

TABLE II
COMPARISON OF EXISTING RESEARCH PAPERS

Author	Year	Technique Used	Dataset	Application	Advantages	Limitations
Sun et al.	2026	Generative AI, Multi-modal Learning	Multimodal Healthcare Dataset	Lesion Localization and Report Generation	Improved disease localization and context-aware report generation	Very high computational and memory requirements
Vavekanand et al.	2026	Multimodal Machine Learning	Predictive Healthcare Dataset	Disease Prediction and Healthcare Analytics	Better prediction accuracy using multiple healthcare data sources	Complex preprocessing and data synchronization
Liu et al.	2025	Deep Learning, Transformer Models	MIMIC-CXR Dataset	Automatic Medical Report Generation	Improved semantic understanding and report quality	Limited explainability and expensive training
Kulsoom et al.	2025	BERT, Transformer-based NLP	Clinical Text Dataset	Radiology Report Classification	Better contextual understanding and classification accuracy	Requires high memory and long training time
Hansen et al.	2024	Multimodal Representation Learning	Healthcare Analytics Dataset	Predictive Healthcare Analytics	Combines imaging, text, and patient history for accurate diagnosis	Difficult integration of heterogeneous healthcare data
Pang et al.	2023	CNN, RNN, Attention Mechanisms	IU X-ray Dataset	Radiology Report Generation	Accurate report generation with attention-based feature extraction	Needs large annotated datasets and high GPU resources
Casey et al.	2021	NLP and Machine Learning	Radiology Reports	Clinical Text Analysis	Efficient extraction of structured clinical information	Difficulty in handling complex medical terminology
Shurrab et al.	2021	Self-Supervised Learning	Medical Imaging Dataset	Medical Image Analysis	Reduces dependency on labeled datasets and improves feature learning	Complex pretraining process and tuning
Van der Velden et al.	2021	Explainable AI (XAI)	Medical Imaging Data	Explainable Medical Diagnosis	Improves trust and transparency in AI healthcare systems	Additional explainability methods increase processing overhead
Messina et al.	2020	Deep Learning with Explainability	Medical Imaging Reports	Automatic Medical Report Generation	Enhances interpretability of generated reports	Explainability may reduce model efficiency and speed

C. Clinical Decision Support

AI-assisted diagnosis and treatment decisions are made through the analysis of clinical data.

D. Medical Image Analysis

CNN algorithms can analyze images to detect abnormality and enhance diagnostic efficiency.

E. Healthcare Recommendation Systems

ML-based recommendation systems make personalized healthcare recommendations for patients.

VII. CHALLENGES IN MEDICAL REPORT ANALYSIS

Despite significant advancements, several challenges remain in ML-based healthcare systems.

A. Data Privacy and Security

Healthcare data consists of critical personal details of patients, which need to be protected against any unauthorized access. Data privacy in these scenarios is a challenging task.

B. Lack of Large Annotated Datasets

Doctors commonly use annotations and abbreviations in the terms used. Healthcare data sets are insufficient and require annotations by experts.

C. Imbalanced Data

Doctors commonly use annotations and abbreviations in the terms used. Healthcare data sets are insufficient and require annotations by experts.

D. Explainability

Advanced deep learning algorithms act as black boxes; thus, healthcare practitioners do not always trust them.

E. Computational Complexity

DL algorithms need highly powerful computational resources.

VIII. FUTURE SCOPE

Future research directions in medical report analysis include:

- Explainable AI systems for trustworthy healthcare
- Federated learning for secure medical data analysis
- Real-time hospital integration systems
- Multimodal AI systems combining text and images
- Large Language Models (LLMs) for medical NLP
- Personalized healthcare recommendation systems
- Lightweight deep learning models for mobile healthcare applications

The integration of AI with healthcare systems is expected to significantly improve diagnosis accuracy and healthcare efficiency in the future.

IX. CONCLUSION

Using machine learning and deep learning techniques, intelligent prediction of diseases, clinical text analysis, and automatic report generation have been achieved. This literature review article focuses on research papers published in the years 2021–2026 regarding the application of machine learning algorithms in health care systems. Machine learning and deep learning algorithms such as SVM, Random Forest, CNN, RNN, LSTM, and Transformer models have been discussed, along with their applications, benefits, and drawbacks. There are some notable challenges such as data privacy, inadequate data sets, lack of explainability, imbalance in data sets, and computation. Despite the presence of these challenges, there is still room for future development of intelligent health care systems in terms of multimodal learning, explainable AI, and generative AI. Intelligent analysis systems of machine learning-assisted medical reports can help in increasing the efficiency of health care services, reducing mistakes made during diagnosis, and aiding healthcare professionals in decision making.

REFERENCES

- [1] K. Sun et al., "Review of generative AI for lesion localization and automatic report generation," *Med-X*, vol. 4, no. 5, 2026.
- [2] R. Vavekanand et al., "Multimodal Machine Learning Approaches in Predictive Healthcare Analytics: A Comprehensive Survey," *Archives of Computational Methods in Engineering*, 2026.
- [3] X. Liu, J. Xin, Q. Shen, Z. Huang, and Z. Wang, "Automatic medical report generation based on deep learning: A state of the art survey," *Computerized Medical Imaging and Graphics*, vol. 120, 2025.
- [4] U. Kulsoom, F. Glavin, and M. Bendeche, "Natural Language Processing and Machine Learning for Analysis of Radiology Reports - A Systematic Review," *IEEE Access*, 2025.
- [5] E. Hansen, T. Sagi, and K. Hose, "Multimodal representation learning for medical analytics - a systematic literature review," *Health Informatics Journal*, 2024.
- [6] T. Pang, P. Li, and L. Zhao, "A survey on automatic generation of medical imaging reports based on deep learning," *BioMedical Engineering OnLine*, vol. 22, no. 48, 2023.
- [7] A. Casey et al., "A systematic review of natural language processing applied to radiology reports," *BMC Medical Informatics and Decision Making*, vol. 21, no. 179, 2021.
- [8] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *arXiv preprint arXiv:2109.08685*, 2021.
- [9] B. van der Velden et al., "Explainable artificial intelligence in deep learning-based medical image analysis," *arXiv preprint arXiv:2107.10912*, 2021.
- [10] P. Messina et al., "A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images," *arXiv preprint arXiv:2010.10563*, 2020.