Leveraging Transformer-based Models for Classification and Large Language Models for Information Extraction from Medical Claims Documents

Dian Prambini[1], Kusworo Adi[2], Komang Budi Aryasa[3]
[123]Doctoral Program of Information System
[123]Diponegoro University, Semarang, Indonesia
[1]dianprambini@students.undip.ac.id, [2]kusworoadi@lecturer.undip.ac.id,
[3]komang@telkom.co.id

Abstract— The study explores the use of advanced AI models for document classification and information extraction in Indonesian medical claims documents. Transformer-based models like Layout LMv3 show good performance with an F1 score above 93% for classification. LLMs like Qwen2.5-VL 7B yield a high F1 score (92-97%) for printed documents but face challenges in processing handwritten medical resumes. The methodology involves extensive data preprocessing and hyperparameter optimization to improve model performance. The research emphasizes the importance of balanced data sampling, valid ground truth data, and careful model selection for accuracy and efficiency. Future studies should integrate deep learning techniques and use pre-trained or fine-tuned BERT models for biomedical and clinical domains. These findings support the potential of transformer-based models and LLMs to automate and simplify document processing workflows in health administration.

Keywords— healthcare AI, document classification, information extraction, Transformer, Large Language Model

I. INTRODUCTION

The growing demand for efficient document processing is forcing the healthcare services industry to undergo a significant transformation. Traditional document processing methods are time-consuming and susceptible to human errors [1]. These inefficiencies lead to high-cost spending [2]. A study in the USA found that manual and traditional methods of processing health documents lead to healthcare providers spending up to 40% of their time on administrative tasks [3]. This leads to burnout [4], [5] and decreased job satisfaction [4]. Traditional claim processing delays reimbursement of patient care costs, and the health insurance claim rejection rate reaches 22% due to manual entry errors [3].

Challenges in health document processing drive automation efforts, with technological advancements like AI enabling widespread adoption of this promising technology. AI significantly enhances medical record documentation, reducing administrative and repetitive tasks for medical staff [4]. AI's efficiency improvements in clinical practice documentation systems have been observed, but its reliability in delivering high-quality standards required for medical documents still needs to be enhanced [5].

AI-based technology is also expected to automate medical claim document processing, specifically verification and adjudication processes, by health insurance companies. This automation is carried out to increase process speed, reduce human errors, and improve cost efficiency [6-9]. Medical claim documents often contain unstructured and semi-structured data formats, necessitating the extraction of all relevant data for automatic processing. Besides the complexity and challenges in extracting data from unstructured formats, various AI-based techniques show promising capabilities in performing these tasks, although so far no suitable framework or solution has been found for multimodal unstructured document extraction [10].

This study investigates the use of AI for classifying and extracting medical claim

documents in Indonesia, using a Transformer-based model for classification and LLM (Large Language Model) for information extraction. No study has been published on the digitization and extraction of medical claim data in Indonesia, which consists of multiple documents mostly in unstructured format and presented in hardcopy. This study aims to fill that gap by exploring the effectiveness of AI-based technology in extracting medical claim data in Indonesia, addressing challenges due to the lack of standardization of claim documents among hospitals. To limit the scope, only medical claim documents from private health insurance participants are being examined.

II. LITERATURE REVIEW

Unstructured data processing techniques have evolved rapidly from traditional methods to AI-based technologies, with Transformer-based models and LLMs being explored for understanding context and patterns in text and images. Despite facing challenges like bias and data quality, the use of large and diverse datasets is crucial in this evolution [10], [11]. Challenges identified include template reliance, data quality issues, complex layouts, and dataset limitations. Future development focuses on creating scalable solutions, ensuring the availability of high-quality datasets, and improving validation methods for automation accuracy across industries [11].

Transformer-based models such as BERT and GPT have the potential to enhance NLP in healthcare to improve the accuracy of entity recognition, data classification, and clinical record analysis [12]. The study demonstrates that the use of AI and NLP in automating clinical document classification found that combining several classification models through ensemble techniques and data augmentation significantly improved performance [13]. Transformer-based models have been found to be quite effective in classifying long documents, outperforming conventional methods [14]. A study comparing Transformer-based models for long document classification reveals that simple models like truncated BERT often perform better than advanced models. Moreover, advanced models tend to require significant computational resources, emphasizing the importance of balancing performance and efficiency in document processing [15]. Another study demonstrates that BERT is capable of identifying diagnoses of cognitive decline from unstructured clinical notes with high levels of accuracy, sensitivity, and specificity [16]. AI technology is gaining traction in the healthcare sector with an emphasis on diagnosis, treatment, patient engagement, and administrative workflows [17].

AI is utilized in creating large-scale transformer-based language models for biomedical and clinical tasks, including concept extraction, relation extraction, semantic similarity, natural language reasoning, and medical question answering [18], [19]. LLM's application in biomedical and clinical NLP shows superior performance in generating synthetic clinical records, offering a promising solution for privacy-preserving data sharing and clinical text analysis [19]. The evaluation of various LLM models for classification and extraction tasks on synthetic records and EMR consistently shows high accuracy, often exceeding 98%, in entity extraction and binary classification tasks [20]. Study on InstructGPT, based on GPT-3, demonstrated a high accuracy rate of up to 97% in extracting clinical information from unstructured PDF text. The model's flexibility, ease of use, and effectiveness in information extraction without specific training surpass traditional NLP tools [21]. The combination of LLM and RAG (retrieval-augmented generation) can enhance summarization accuracy and extract key clinical data from unstructured clinical notes [22].

Data extraction from handwritten medical records poses a unique challenge due to their diverse variations, necessitating distinct algorithms and models for recognizing different languages to achieve high accuracy [23]. Studies on handwriting text recognition often focus on deep learning architectures

such as RNN (particularly LSTM and GRU units), CNN, including CRNN, or combinations of these models [10], [24-28]. CNNs and RNNs work together to accurately recognize complex and cursive handwriting by extracting spatial features from images and modeling sequential dependencies in data [10]. Developing hybrid algorithms among deep learning architectures is crucial for getting the best accuracy outcomes [27]. Combining convolutional layers with recurrent units is effective to enhance recognition robustness [29]. Integrating LLM for post-processing correction of handwritten text recognition resulting from deep learning algorithms has also been reported to improve recognition accuracy [30].

The application of LLMs in healthcare requires further analysis of broader impacts and challenges, such as ethics, bias, hallucinations, incomplete data, validation limitations, and variability of results, as well as data and model security [19], [20], [22]. Future focus will be on developing high-quality datasets, more general and adaptive AI models, and multimodal solutions for processing diverse data types simultaneously. Template-free approaches, full automation, and the integration of the latest technologies are expected to accelerate digital transformation in various sectors such as healthcare, finance, and governance [10].

III. METHOD

A. DATASET USED FOR THE STUDY

The study utilized scanned image data from 6,987 real medical claims from hundreds of Indonesian hospitals. The documents are classified based on their layout/shape/content form, text within images, and data extraction requirements. The initial classification process should be conducted without overlap, preventing the use of similar documents in multiple classifications to avoid biased results and maintain consistency. The classification results revealed 41 distinct types of document image classifications to be utilized as the dataset (Figure 1).

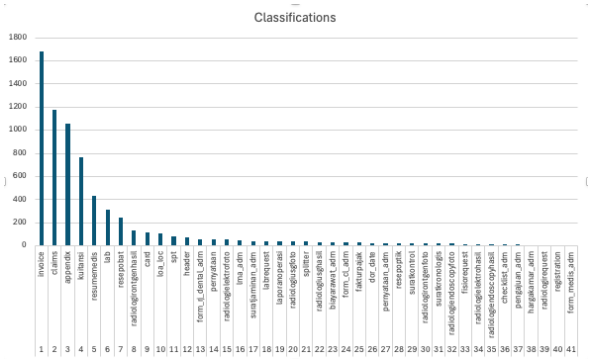

Figure 1. Initial Dataset

B. DATA PRE-PROCESSING PROCEDURE

The data preprocessing procedure that was used is as follows: (1) The augmentation technique: involves modifying the original document to incorporate image variation in each classification, such as rotation, scaling, color adjustment, and/or noise; (2) Handling imbalance datasets: the initial dataset in Figure 1 is imbalanced, requiring oversampling or downsampling to ensure balance in each class. Oversampling adds data from the minority class to improve model learning, while downsampling reduces data from the majority class to avoid bias. The balancing results with a dataset of 10,778; (3) Data transformation: (i) scaling and normalization, which stabilizes data values and improve learning effectiveness by adjusting the data to a range between 0 and 1; (ii) encoding categorical variables (binary code), which converts labels into binary code to reduce data dimensions.

C. THE AI ALGORITHMS FOR DATA CLASSIFICATION AND EXTRACTION

A BERT-based model for document comprehension tasks serves as the foundation for classification. This model takes words that are converted into tokens as input, which are then converted into numbers and processed in multilayered neural networks. That process is known as the BERT Masked Language Model. Details regarding the document's layout or shape are also necessary for classification. Layout processing involves examining the word's location (x-y

coordinates). In addition to position, word image features such as bold, italic, and so forth are required to enhance word visualization. The masked image modeling task is used to extract the image picture. In general, this architecture combines text, layout, and visual representations to enhance the overall understanding of documents [31], [32] which has proven to improve classification performance [31]. The model suitable for this classification task is the Transformer-based Layout LM (Layout Language Model), and classification model in this study using Layout LMv3. Layout LM models are designed to handle scanned documents with complex layouts such as tables and forms [33].

Initially, the data extraction model uses the same algorithm as the classification model, which adds labels to the documents by conducting annotation or labeling positions (x-y coordinates) with the labels required to generate the dataset. This modeling requires more time and effort to perform labeling on all documents. The extraction results are quite good for position determination, but the extracted text results are less accurate. Given those findings, using a multimodal LLM such as Qwen that can comprehend and combine various kinds of data is necessary to manage the intricacy of medical data [34]. The model utilized in this investigation is Qwen2.5-VL 7B, a sophisticated Large Vision Language Model (LVLM) variant of the Qwen family that integrates language and visual abilities into a single large model with promising medical application prospects [35]. Qwen2.5-VL is a powerful tool that accurately localizes objects, extracts structured data from various sources, and provides detailed analysis of charts, diagrams, and layouts [36]. However, the use of LLM requires a fairly large computing machine. Qwen balances efficiency and resource utilization in LLMs, recommended for compliance sector workers without high-end computational resources [37]. The hardware used in this study has the following spesifications: (1) Classification using 12 vCPU, 170 GB Memory, GPU 1 x NVIDIA A100 80 GB; (2) Extraction using 16 vCPU, 32 GB Memory, GPU 1 x NVIDIA T4.

IV. RESULTS AND DISCUSSION

A. RESULTS OF THE CLASSIFICATION MODEL

The dataset consisting of 10,778 data points is divided into training data and validation data, with each portion being 80%:20% or 70%:30%. Here are some hyperparameters used in the training process: Number of training epochs = 10; batch size = 5; learning rate = 5e-6; weight decay = 0.01; best model metric is F1. The number of epochs refers to how many times the entire training dataset is fully processed by the model. The model's generalization on new data requires the right number of epochs, as too few can cause underfitting and too many can lead to overfitting. Figure 2 shows the effect of the number of epochs (horizontal axis) on the model performance metrics.

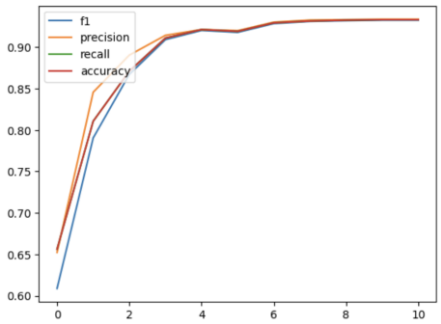

Figure 2. The number of epoch (horizontal axis) Vs the model performance metrics

Batch size is a measure of how much data is processed simultaneously in one iteration before updating model parameters; the larger the batch size, the more stable and faster the training process becomes, but it also requires more memory. The learning rate is a crucial hyperparameter that determines the model's weight update step during training iterations, and the correct learning rate value is essential for achieving an optimal and stable model. Weight decay is a regularization technique in machine learning that prevents overfitting by gradually reducing weights during training, thereby maintaining simplicity and facilitating generalization to new data. The F1 metric model is the most suitable for

imbalanced classification tasks, as it provides a balanced picture of the model's precision in accurately identifying relevant information (precision) and its completeness in capturing all critical data (recall), ensuring a more accurate representation of the model's performance.

B. DISCUSSION OF THE CLASSIFICATION MODEL

Appropriate evaluation metrics are crucial for selecting the best classifier model during classification training, as they help discriminate and acquire the most suitable model [38]. The confusion matrix is a tool used to measure the detailed performance of a model by comparing the number of correct and incorrect predictions to the actual labels (ground truth). It helps identify how the model makes mistakes and calculates metrics such as accuracy, precision, recall, and F1-score. Accuracy measures the number of correct predictions, precision calculates the number of all predicted positives, recall measures the number of actual positives correctly identified by the model, and the F1 score is the harmonic mean of precision and recall. The confusion matrix can identify bias in a model, indicating if it misclassifies certain classes more frequently, enabling the identification of dataset classifications that need improvement for more accurate predictions.

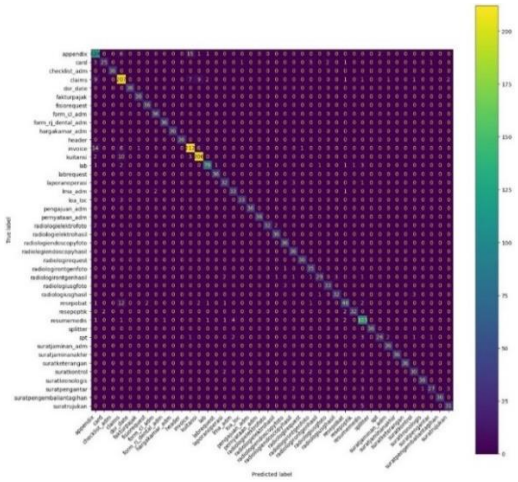

Figure 3. Confusion Matrix of Classification Model Training Scenario 6

This study analyzed six model training scenarios, varying in proportion between training and validation data and the presence or absence of oversampling or downsampling. Performance was measured in terms of training loss, validation loss, accuracy, precision, recall, and F1-score. Training loss measures the model's error against the training dataset, while validation loss measures the model's error on new data, assessing its generalization ability. This metrics are crucial in evaluating the model's performance in real-world scenarios. Figure 3 shows the confusion matrix of the classification training model scenario 6, with the vertical axis representing the true labels and the horizontal axis representing the predicted labels. From this matrix, one can observe how the model classifies each class.

Table 1 presents six model training scenarios and their performance measurement results. While scenarios 1 and 2 show no significant difference, some performance metrics slightly improve, suggesting no significant impact of downsampling. The outcomes of scenarios 1 and 3 do not differ significantly when the ratios of training and validation data are altered. The difference becomes significant when oversampling data is increased, as seen in scenarios 4 and 5. The best result was achieved when the selection of oversampling and downsampling sizes was just right, resulting in all performance metrics exceeding 93%.

Table 1. Performance Metrics of 6 Classification Model Training Scenarios

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
Data Train	7814 – 80%	6602 – 80%	6837 – 70%	7544 – 70%	8622 – 80%	7410 – 80%
Data Valid	1954 – 20%	1651 – 20%	2931 – 30%	3234 – 30%	2156 – 20%	1853 – 20%
Over Sampling	120	120	120	150	150	150
Down Sampling	-	800	-	-	-	800
Epoch	10	10	10	10	10	10
Training Loss	0.1281	0.1864	0.1699	0.1196	0.0904	0.115300
Validation Loss	0.492001	0.413155	0.463603	0.370335	0.349696	0.312154
Accuracy	0.903787	0.906723	0.908222	0.922078	0.926252	0.933081
Precision	0.904560	0.910386	0.909442	0.922681	0.926646	0.933632
Recall	0.903787	0.906723	0.908222	0.922078	0.926252	0.933081
F1 Score	0.902828	0.905892	0.907434	0.921405	0.92593	0.932597

The results of scenario 6 show a decrease in validation loss but slightly increased training loss, which is less common but can

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v12i2.14285

occur due to various factors. One possibility is regularization, which increases training loss but helps avoid overfitting and perform better on new data. Another possibility is a small batch size causing fluctuating training loss between epochs, while decreasing validation loss indicates overall improvement. A high learning rate can also cause fluctuating training loss. Table 1 reveals that validation loss is generally higher than training loss, possibly due to insufficient training data, unequal distribution of training and validation data, inconsistent data augmentation, ambiguous validation data due to noise or incorrect labels, or suboptimal hyperparameter tuning.

Table 2. Comparison of Performance Metrics for Layout LM and ViT

Parameter	Layout LMv3 (Scenario 5)	ViT (Scenario 5)	Layout LMv3 (Scenario 6)	ViT (Scenario 6)
Data Train	8622 – 80%	8622 – 80%	7410 – 80%	7410 – 80%
Data Valid	2156 – 20%	2156 – 20%	1853 – 20%	1853 – 20%
Over Sampling	150	150	150	150
Down Sampling	-	-	800	800
Epoch	10	10	10	10
Training Loss	0.0904	0.743600	0.115300	0.774900
Validation Loss	0.349696	0.865608	0.312154	0.886144
Accuracy	0.926252	0.751391	0.933081	0.752700
Precision	0.926646	0.759294	0.933632	0.760602
Recall	0.926252	0.751391	0.933081	0.752500
F1 Score	0.92593	0.736214	0.932597	0.746130

The study compares the performance of another Transformer-based model, ViT (Vision Transformer), a deep learning architecture for image classification and vision-related tasks. ViT divides images into small patches, generates patch embedding, adds class tokens for representation and positional encoding to determine their order, then processes them using a transformer layer (encoder). The class tokens that come out of the encoder are used to retrieve the label, which will be used for the final class prediction by passing the class token values to MLP (multilayer perceptron) head [39], [40]. A study demonstrates the effectiveness of ViT models in medical image analysis by capturing global dependencies and detailed features for image classification [41]. The comparison was conducted in training scenarios 5 and 6, with the results shown in Table 2. Overall results show that the model using LayoutLM outperforms the model using ViT.

C. RESULTS OF THE EXTRACTION MODEL

The extraction process utilizes Qwen2.5-VL 7B, which has an OCR pipeline, to read and understand text from images. Specific prompt engineering is employed to identify and extract crucial information from documents, ensuring optimal output [42].

The study extracted information from five types of classified documents: medical resumes, LOA (letters of acceptance), claim documents, invoices, and receipts. The type and amount of information needed for subsequent claim verification and adjudication were determined for each document. A ground truth table was created to calculate the model's performance metrics. The results for the samples used are presented in Table 3.

Table 3. Performance Metrics of the Extraction Model for 5 Types of Documents

Parameter	Medical Resume	LOA	Claim	Invoice	Receipt
No. of Document	75	78	50	50	50
No. Information/Doc	2	2	8	3	3
Total Information	150	156	400	150	150
False Negative (FN)	99	21	58	9	10
True Positive (TP)	50	135	342	141	140
True Negative (TN)	0	0	0	0	0
False Positive (FP)	1	0	0	0	0
Accuracy	33.33%	86.53%	85.50%	94.00%	93.33%
Precision	98.04%	100.00%	100.00%	100.00%	100.00%
Recall	33.56%	86.54%	85.50%	94.00%	93.33%
F1 Score	50.00%	92.78%	92.18%	96.91%	96.55%

D. DISCUSSION OF THE EXTRACTION MODEL

The precision metric shows the highest value among performance metrics, indicating that most positive predictions were correct. Positive prediction errors were only observed in medical resume extraction. Accuracy and recall for invoice and receipt documents are above 93% and the F1-score is above 96%, respectively. LOA and claim documents have lower accuracy and recall but are still above 85%, and the F1 score is lower but still above 92%. Medical resume documents have the lowest results at 33.33%, 33.56%, and 50.00%, respectively. The low performance metrics may be due to the model not being optimal for extracting information from handwritten notes, as only medical resumes in the form of handwritten notes are used in the sample.

The model's performance in extracting information from printed documents is satisfactory, with an F1-score above 92% and even exceeding 96% for some documents. This indicates that AI can automate the processing of medical claim documents. However, the main challenge is improving the model's performance in extracting information from handwritten notes. A comprehensive evaluation of the model's performance metrics and a larger sample size is needed to determine appropriate improvement efforts. The model's initial evaluation provides optimism for AI's potential in medical claim document processing.

V. CONCLUSION

The study explores the potential of AI in automating medical claim document processing in Indonesia, where there is a lack of standardized medical claim documents and most are in hardcopy form. The LayoutLMv3 model, a transformer-based AI architecture, shows good results in document classification with an F1-score above 93%. Compared to ViT, another transformer-based algorithm, it has a lower F1-score of 74%. The LLM Qwen2.5-VL 7B model also shows good results for data extraction with an F1 score of 92-97% for printed documents but less satisfactory results for handwritten notes at 50%.

To improve the model's learning capabilities, more extensive and diverse datasets must be used in future research. Finding reliable and unambiguous ground truth data for classification and extraction tasks is one of the challenges. Combining deep learning algorithms like CNN, RNN, and CRNN to improve handwritten note extraction performance along with the use of pre-trained or optimized BERT models designed for the biomedical and/or clinical domain, like BioBERT, ClinicalBERT, PubMedBERT, and GatorTron, to boost the extraction model's overall performance is another intriguing research topic.

REFERENCES

[1] R. Saxena, G. Katage, C. Kumar, N. M. Pathan, and M. N. Bargir, “AI redefining healthcare documentation for tomorrow: Exploring the impact of ai on healthcare documentation,” in Computational Convergence and Interoperability in Electronic Health Records (EHR), IGI Global, 2024, pp. 51–66. doi: 10.4018/979-8-3693-3989-3.ch003.

[2] M. Pandey, M. Arora, S. Arora, C. Goyal, V. K. Gera, and H. Yadav, “AI-based Integrated Approach for the Development of Intelligent Document Management System (IDMS),” in Procedia Computer Science, 2023, pp. 725–736. doi: 10.1016/j.procs.2023.12.127.

[3] Ramesh Pingili, “AI-driven intelligent document processing for healthcare and insurance,” International Journal of Science and Research Archive, vol. 14, no. 1, pp. 1063–1077, Jan. 2025, doi: 10.30574/ijsra.2025.14.1.0194.

[4] A. R. Bongurala, D. Save, A. Virmani, and R. Kashyap, “Transforming Health Care with Artificial Intelligence: Redefining Medical Documentation,” Sep. 01, 2024, Elsevier B.V. doi: 10.1016/j.mcpdig.2024.05.006.

[5] A. Bracken, C. Reilly, A. Feeley, E. Sheehan, K. Merghani, and I. Feeley, “Artificial Intelligence (AI) – Powered Documentation Systems in Healthcare: A Systematic Review,” Dec. 01, 2025, Springer. doi: 10.1007/s10916-025-02157-4.

[6] H. Kong Journal and M. B. H. Kong, “Revolutionizing Claims Processing in the Healthcare Industry: The Expanding Role of Automation and AI.”

[7] G. Vaithiyalingam, “Bridging the Gap: AI, Automation, and the Future of Seamless Healthcare Claims Processing.”

[8] A. Siddiqui and L. Boukhalfa, “Streamlining Healthcare Claims Processing Through Automation: Reducing Costs and Improving Administrative Workflows,” AI-Assisted

Scientific Discovery by Science Academic Press.

[9] J. Hernandez, "African Journal of Artificial Intelligence and Sustainable Development Volume 4 Issue 1 Semi Annual Edition | Advancing Healthcare Claims Processing with Automation: Enhancing Patient Outcomes and Administrative Efficiency African Journal of Artificial Intelligence and Sustainable Development Volume 4 Issue 1 Semi Annual Edition | African Journal of Artificial Intelligence and Sustainable Development Volume 4 Issue 1 Semi Annual Edition |."

[10] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," J Big Data, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00948-z.

[11] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, "Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions," 2021, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ACCESS.2021.3072900.

[12] V. P. Vasani, S. C. Pawar, P. Sowmya, S. Ahamad, A. Sahu, and G. Talele, "Transformer Models for Enhanced Natural Language Processing in Medical Records Management," in Proceedings - 4th International Conference on Technological Advancements in Computational Sciences, ICTACS 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1808–1814. doi: 10.1109/ICTACS62700.2024.10840744.

[13] M. Youssef, M. Abu-Elkheir, and M. Mashaly, "Automating Clinical Document Classification: AI Solutions for Enhanced Healthcare Decision Support," in NILES 2024 - 6th Novel Intelligent and Leading Emerging Sciences Conference, Proceedings, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 345–348. doi: 10.1109/NILES63360.2024.10753219.

[14] X. Dai, I. Chalkidis, S. Darkner, and D. Elliott, "Revisiting Transformer-based Models for Long Document Classification," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2204.06683

[15] H. H. Park, Y. Vyas, and K. Shah, "Efficient Classification of Long Documents Using Transformers," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.11258

[16] R. Shankar, A. Bundele, and A. Mukhopadhyay, "Natural language processing of electronic health records for early detection of cognitive decline: a systematic review," NPJ Digit Med, vol. 8, no. 1, Dec. 2025, doi: 10.1038/s41746-025-01527-z.

[17] T. Davenport and R. Kalakota, "DIGITAL TECHNOLOGY The potential for artificial intelligence in healthcare," 2019.

[18] X. Yang et al., "A large language model for electronic health records," NPJ Digit Med, vol. 5, no. 1, Dec. 2022, doi: 10.1038/s41746-022-00742-2.

[19] C. Peng et al., "A study of generative large language model for medical research and healthcare," NPJ Digit Med, vol. 6, no. 1, Dec. 2023, doi: 10.1038/s41746-023-00958-w.

[20] V. Ntinopoulos et al., "Large language models for data extraction from unstructured and semi-structured electronic health records: A multiple model performance evaluation," BMJ Health Care Inform, vol. 32, no. 1, Jan. 2025, doi: 10.1136/bmjhci-2024-101139.

[21] V. Sciannameo et al., "Information extraction from medical case reports using OpenAI InstructGPT," Comput Methods Programs Biomed, vol. 255, Oct. 2024, doi: 10.1016/j.cmpb.2024.108326.

[22] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from

electronic health records," J Biomed Inform, vol. 156, Aug. 2024, doi: 10.1016/j.jbi.2024.104662.

[23] K. Saini, K. Sharma, A. Agarwal, K. Jayan, and D. Dev, "Handwritten Text Recognition Using Machine Learning," in 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology, ICSEIET 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 121–124. doi: 10.1109/ICSEIET58677.2023.10303304

[24] D. Gifu, "AI-backed OCR in Healthcare," in Procedia Computer Science, Elsevier B.V., 2022, pp. 1134–1143. doi: 10.1016/j.procs.2022.09.169.

[25] L. Navya, M. F. Ali, K. P. Sai, K. Shyam, and A. Ramesh, "Handwritten Text Recognition Using Deep Learning Techniques," in 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems, AICERA/ICIS 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/AICERA/ICIS59538.2023.10420040.

[26] M. J. Hossain, S. Samiha Zaman, F. R. Akash, M. Rifat Sarker, and M. R. Huq, "Developing a Bangla Handwritten Text Recognition Framework using Deep Learning," in 2023 26th International Conference on Computer and Information Technology, ICCIT 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCIT60459.2023.10441107.

[27] V. Jayanthi and S. Thenmalar, "Handwritten Word Recognition of Various Languages: A Review," in 9th International Conference on Smart Computing and Communications: Intelligent Technologies and Applications, ICSCC 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 524–528. doi: 10.1109/ICSCC59169.2023.10334936.

[28] S. Sinha, G. Paliwal, and K. P. Sharma, "Advanced Handwritten Text Recognition and Analysis System," in 2025 Global Conference in Emerging Technology (GINOTECH), IEEE, May 2025, pp. 1–8. doi: 10.1109/GINOTECH63460.2025.11076990.

[29] R.S. Srichandra, P. S. Rahul, M. S. Govind, V. Battula, « Performance of Convolutional Recurrent Networks for Handwritten Text Recognition », 10th International Conference on Computing for Sustainable Global Development. New Delhi, India, pp. 1207-1210 IEEE, 2023.

[30] M. P. Kalra, A. Kushwaha, and P. P. Vuppuluri, "LLM Powered HTR: Integrating Handwritten Text Recognition System with Large Language Model," in 2024 IEEE Students Conference on Engineering and Systems: Interdisciplinary Technologies for Sustainable Future, SCES 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/SCES61914.2024.10652342.

[31] K. Fathima, A. Athar, and H.-C. Kim, "Multi-Class Document Classification using LayoutLMv1 and V2."

[32] U. I. Avei, D. Goularas, E. E. Korkmaz, and B. Deveci, "Information Extraction from Scanned Invoice Documents Using Deep Learning Methods," in Proceedings - 13th International Conference on Image Processing Theory, Tools and Applications, IPTA 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IPTA62886.2024.10755641.

[33] M. S. Baysan, F. Kizilay, A. I. Özmen, and G. Ince, "Document Classification and Key Information Extraction Using Multimodal Transformers," in UBMK 2024 - Proceedings: 9th International Conference on Computer Science and Engineering, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 276–281. doi: 10.1109/UBMK63289.2024.10773451.

[34] B. Wang et al., "DeepSeek and Qwen in Healthcare: Pioneering Multimodal Large Language Models for Nextgeneration Disease Prediction," in 2025 10th International Conference on Computer and Communication System (ICCCS), IEEE, Apr. 2025, pp. 100–105. doi:10.1109/ICCCS65393.2025.11069555.

[35] J. Luo, H. Yu, C. Tan, and H. Yu, "Enhanced Qwen-VL 7B Model via Instruction Finetuning on Chinese Medical Dataset," in 2024 5th International Conference on Computer Engineering and Application, ICCEA 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 526–530. doi: 10.1109/ICCEA62105.2024.10603938.

[36] S. Bai et al., "Qwen2.5-VL Technical Report," Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.13923

[37] R. Phogat, D. Arora, P. S. Mehra, J. Sharma, and D. Chawla, "A Comparative Study of Large Language Models: ChatGPT, DeepSeek, Claude and Qwen," in 3rd IEEE International Conference on Device Intelligence, Computing and Communication Technologies, DICCT 2025, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 609–613. doi: 10.1109/DICCT64131.2025.10986449.

[38] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[39] M. S. Mia, A. B. H. Arnob, A. Naim, A. A. B. Voban, and M. S. Islam, "ViTs are Everywhere: A Comprehensive Study Showcasing Vision Transformers in Different Domain," in 2023 International Conference on Cognitive Computing and Complex Data, ICCD 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 101–117. doi: 10.1109/ICCD59681.2023.10420683.

[40] D. Pantelaios, P. A. Theofilou, P. Tzouveli, and S. Kollias, "Hybrid CNN-ViT Models for Medical Image Classification," in Proceedings - International Symposium on Biomedical Imaging, IEEE Computer Society, 2024. doi: 10.1109/ISBI56570.2024.10635205.

[41] M. G. Dahmani, M. Tarhouni, and S. Zidi, "Vision Transformers (ViT) for Enhanced Skin Cancer Classification," in 2024 IEEE International Conference on Artificial Intelligence and Green Energy, ICAIGE 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICAIGE62696.2024.10776698.

[42] Y. Yao, Z. Lin, X. Liu, and Y. Li, "Document Information Extraction in Engineering Reports Through Prompt Engineering with Large Language Models," Institute of Electrical and Electronics Engineers (IEEE), Jun. 2025, pp. 1–4. doi: 10.1109/ainit65432.2025.11035034.