scientific reports

OPEN

Enhanced effective convolutional attention network with squeeze-and-excitation inception module for multi-label clinical document classification

M. Venkata Krishna Reddy[1]✉, L. Raghavendar Raju[2], Kashi Sai Prasad[3], Dr. D Anitha Kumari[4], Vadlamani Veerabhadram[5] & Nagendar Yamsani[6]

Clinical Document Classification (CDC) is crucial in healthcare for organizing and categorizing large volumes of medical information, leading to improved patient care, streamlined research, and enhanced administrative efficiency. With the advancement of artificial intelligence, automatic CDC is now achievable through deep learning techniques. While existing research has shown promising results, more effective and accurate classification of long clinical documents is still desired. To address this, we propose a new model called the Enhanced Effective Convolutional Attention Network (EECAN), which incorporates a Squeeze-and-Excitation (SE) Inception module to improve feature representation by adaptively recalibrating channel-wise feature responses. This architecture introduces an Encoder and Attention-Based Clinical Document Classification (EAB-CDC) strategy, which utilizes sum-pooling and multi-layer attention mechanisms to extract salient features from clinical document representations. This study proposes EECAN (Enhanced Effective Convolutional Attention Network) as the overall model architecture and EAB-CDC (Encoder and Attention-Based Clinical Document Classification) as a core strategy conducted in EECAN. EAB-CDC is not a standalone model but a functional part applied to the architecture for discriminative feature extraction by sum-pooling and multi-layer attention mechanisms. With this integrated design, EECAN can transform multi-label clinical texts' general and label-specific contexts without losing information. Our empirical study, conducted on benchmark datasets such as MIMIC-III and MIMIC-III-50, demonstrates that the proposed EECAN model outperforms several existing deep learning approaches, achieving AUC scores of 99.70% and 99.80% using sum-pooling and multi-layer attention, respectively. These results highlight the model's substantial potential for integration into clinical systems, such as Electronic Health Record (EHR) platforms, for the automated classification of clinical texts and improved healthcare decision-making support.

Keywords Clinical document classification, Convolutional attention network, Artificial intelligence, Deep learning, Squeeze and excitation inception

Clinical Document Classification (CDC) is essential in healthcare for efficiently organizing and extracting meaningful information from large volumes of unstructured medical text. By categorizing documents based on their content, healthcare providers can enhance the speed and accuracy of information retrieval, supporting critical functions such as diagnosis, treatment planning, billing, and clinical decision-making. During the last few years, electronic health records (EHRs) have been abundant in unstructured clinical narratives. Clinical Document Classification (CDC) is a task directly used to structure this valuable information. This task is

[1]Department of Computer Science and Engineering, Chaitanya Bharathi Institute of Technology (Autonomous), Gandipet, Hyderabad, India. [2]Department of Computer Science and Engineering, Matrusri Engineering College, Hyderabad, India. [3]Department of CSE-AI&ML, , MLR Institute of Technology, Hyderabad, India. [4]Professor, Department of CSM, TKR College of Engineering and Technology, Hyderabad, India. [5]CVR College of Engineering, Hyderabad, Telangana, , India. [6]School of Computer Science and Artificial Intelligence, SR University, Warangal, India. ✉email: krishnareddy_cse@cbit.ac.in

challenging due to the complexity of medical language, the co-occurrence of multiple diagnoses, diverse writing styles, and long unstructured text. The concept of CDC becomes increasingly complicated in multi-label scenarios, where a document can be relevant to various medical codes or conditions. However, standard machine learning and rule-based systems generalize poorly between institutions and struggle with the heterogeneity and sparsity of clinical datasets, necessitating more robust deep learning-based approaches.

Despite advances in natural language processing (NLP) and deep learning, CDC remains challenging—particularly for long clinical documents—due to the complexity of medical language, co-occurring conditions, diverse document formats, and redundant or ambiguous information. Various deep learning techniques have been applied to CDC, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), attention mechanisms, and transformer-based models. While these models have demonstrated promising performance, they often face limitations in modeling long-range dependencies, handling label imbalance, and efficiently extracting relevant features from lengthy, multi-label documents. Many existing models struggle to capture fine-grained, label-specific features in long sequences due to vanishing gradients, loss of contextual relevance, and computational constraints.

Although significant progress has been in applying CNNs, RNNs, and transformer-based architectures to clinical NLP, these models have fundamental limitations in accessing long, multi-label clinical documents. So do Transformers (which are powerful but computationally expensive for long sequences) or CNNs and RNNs (which tend to lose context relevance over a text length). Moreover, the current attention mechanisms might fail to discriminate between contexts-specific labels, resulting in potential misclassifications. These case management challenges give rise to a research gap in a practical, scalable, and contextual model needed to accurately classify complex clinical narratives with multiple overlapping conditions.

To address these limitations, we propose an Enhanced Effective Convolutional Attention Network (EECAN), which extends the baseline EffectiveCAN model[1]. EffectiveCAN was chosen due to its demonstrated capability to leverage attention mechanisms for CDC tasks. However, to improve feature representation and classification performance, we enhance the model with a Squeeze-and-Excitation Inception (SE-Inception) module that adaptively recalibrates channel-wise feature responses, enabling the model to focus on informative features. We also introduce an Encoder and Attention-Based Clinical Document Classification (EAB-CDC) strategy within the EECAN framework, which employs sum-pooling and multi-layer attention mechanisms to capture global and label-specific information effectively. EAB-CDC is not a separate model but an integral component of EECAN that enhances its attention-based feature extraction. Our proposed model is evaluated on benchmark datasets, including MIMIC-III and MIMIC-III-50, and demonstrates superior performance compared to existing CNN-based and transformer-based models. EECAN addresses the critical challenges of extended document classification, multi-label learning, and feature sparsity by combining SE-Inception and multi-level attention in a unified architecture.

Our work's novelty is integrating the SE-Inception module with multi-layer and sum-pooling attention mechanisms into a unified CNN-based architecture specifically for multi-label clinical text classification. In contrast to existing frameworks, EECAN employs channel-wise recalibration and hierarchical attention to address the challenges of feature sparsity, label imbalance, and long-range dependencies. Our contributions are three-fold: we propose a novel boost architecture, validate it on various benchmark datasets, and substantially outperform existing CNN- and transformer-based models on these datasets. The key contributions of this work are as follows:

1. We propose a novel architecture, EECAN, which enhances EffectiveCAN with SE-Inception and attention mechanisms to improve feature extraction and classification accuracy for multi-label clinical documents.
2. We introduce the EAB-CDC strategy within EECAN to capture hierarchical and label-specific features through sum-pooling and multi-layer attention.
3. We conduct extensive experiments on multiple benchmark datasets, demonstrating EECAN's superior performance compared to existing state-of-the-art methods.

The remainder of this paper is organized as follows: Section "Related work" presents a review of related work; section "Materials and methods" describes the materials and proposed methodology; section "Experimental results" provides experimental results and analysis; section "Discussion" discusses findings and future directions; and section "Conclusion and future scope" concludes the paper.

Related work

This literature review discusses diverse techniques for machine learning and natural language processing (NLP) for classifying clinical documents. Wang et al.[2] utilized pre-trained embeddings to accelerate machine learning and rule-based NLP. They applied deep representation and inadequate oversight to enable the automation of clinical text categorization labeling. They found that CNNs outperformed other methods but also identified some limitations. Kadhim[3] focused on text classification for structured information extraction, emphasizing the utility of k-NN. Khanday et al.[4] highlighted the impact of artificial intelligence on healthcare, particularly in addressing COVID-19. They proposed that recurrent neural networks may offer promise in this regard. Kim et al.[5] employed multi-co-training (MCT) to enhance performance in document labeling. Cong et al.[6] discussed the precise localization of LINP in biological tissues using X-ray fluorescence computed tomography (XFCT). Chan et al.[7] explored machine learning (ML) applications in dermatology, focusing on improving customized care, mobile app evaluations, and illness categorization. Huang et al.[8] addressed non-IID issues in ICU patient data clustering using Community-based federated machine learning (CBFL). Wu et al.[9] employed a random forest model to predict fatty liver disease accurately. Daghistani et al.[10] created an artificial intelligence model to predict the duration of hospital stays for cardiac patients. Reddy et al.[11] developed an algorithm-based ensemble artificial intelligence model to identify diabetic retinopathy. They suggested the potential for larger datasets in future research.

In a study on heart disease prediction in an IoT and cloud-based healthcare application, Ganesan and Kumar[12] found that J48 classifiers outperformed other models regarding accuracy, precision, and recall. Picciali et al.[13] advanced medical fields by handling large amounts of health-related data with deep learning (DL), but they noted challenges with personalization and ethical issues. Diwakar et al.[14] aimed to save lives by using artificial intelligence for early cardiac illness detection, and they also emphasized the importance of dataset quality in their review of classification techniques. Sarker et al.[15] examined various machine learning algorithms, focusing on their prospective uses in practical contexts such as cyber security and healthcare. Souri et al.[16] presented a machine learning-powered Internet of Things (IoT)-based student healthcare monitoring system that can identify behavioral and physiological changes, with the support vector machine (SVM) classifier being remarkably accurate.

Ebrahimi et al.[17] explored modern deep learning (DL) techniques in medicine, particularly in categorizing ECG signals, highlighting the effectiveness of CNN for feature extraction in achieving good classification accuracy for arrhythmias. Shorten et al.[18] discussed the applications, challenges, and potential solutions related to deep learning's involvement in combating COVID-19 across various fields. Sorin et al.[19] investigated deep learning in NLP for radiology to improve performance and assess relevant literature. Fawaz et al.[20] extensively investigated deep learning for Time Series Classification (TSC), examining current architectures, developing an open-source framework, and establishing a single taxonomy. They also emphasized the need for more research on data augmentation, transfer learning, and normalization effects in TSC investigations. Finally, Battineni[21] found that using support vector machines (SVM) with specific parameters that produced the best results for dementia prediction had a precision of 64.18% and an accuracy of 68.75%. This study underscores the significance of SVM in precise prediction, considering the substantial health risk posed by dementia, especially to the elderly.

Zhang et al.[22] suggested the SDL model to address challenges in medical image classification by utilizing multiple DCNNs. Future work should focus on scaling the model and using reinforcement learning to optimize the number of DCNNs. Wang et al.[23] highlighted the need for better performance with deep learning models, which excel in mammography classification but struggle with external data fluctuations. Navamani[24] utilized neural networks, particularly CNNs, for feature extraction and classification. Deep learning in health informatics takes advantage of rich biological data. Amyar et al.[25] presented a multitask deep learning model for COVID-19 segmentation and identification in chest CT images, outperforming cutting-edge methods in both tasks.

Bohr et al.[26] discussed the increasing demand for healthcare services, especially from physicians, emphasizing the essential nature of technology and smartphone on-demand services. They highlighted that AI in healthcare outperforms humans in various tasks, including diagnosis and therapy. Ali et al.[27] recommended using feature fusion and ensemble deep learning techniques to create an intelligent healthcare system for predicting cardiac disease, providing better diagnostics and efficient risk factor identification. Data mining approaches will be used in future studies to improve feature fusion. Israel et al.[28] investigated the limited use of machine learning in clinical care, examining potential applications, difficulties, and obstacles. Marshall et al.[29] aimed to expedite systematic reviews, emphasizing how contemporary technologies automate data extraction, filtering, and search tasks. They noted that reliable tool maintenance remains an ongoing issue. Zhou et al.[30] examined deep learning techniques for segmenting multi-modal medical images. They highlighted the importance of multi-modality in medical imaging, emphasizing and contrasting the efficacy of fusion techniques.

Solares et al.[31] examined the limited uses of deep learning for clinical decision support but observed an increasing trend of personalized prediction using electronic health records (EHR). Subsequent investigations will explore methods from other fields and apply them to various medical data. Sarker et al.[32] explained the significance of using contextual data to estimate and forecast smartphone usage to create tailored systems aware of context. Based on the efficacy study, future studies may involve creating practical applications for intelligent, tailored services. Houssein et al.[33] explained computer-aided detection and emphasized the need for early breast cancer detection. It discusses convolutional neural networks and classifies methods like SVM, DT, ANN, Naive Bayesian networks, and nearest neighbors. The study focuses on upcoming developments and difficulties in the diagnosis and categorization of breast cancer.

Shamshirband et al.[34] explained how the complexity of healthcare data has led to a growing usage of machine learning in medicine, especially Deep Neural Network (DNN) models. Memory loss and time waste are issues, indicating the necessity for efficient techniques. Future work on Explainable AI for distributed systems, model enhancements, and hardware requirements may highlight the possible uses of deep learning in medical fields. Zhao et al.[35] addressed the issues with privacy in deep learning because of sensitive data. The efficacy of SecProbe is assessed using real-world datasets, demonstrating its resilience against untrustworthy participants and attaining accuracy levels on par with conventional centralized models while guaranteeing strict privacy protection.

Ganggayah et al.[36] applied machine learning methods to a dataset from Malaysia to investigate breast cancer survival markers. Key characteristics such as cancer stage, tumor size, and therapy kind were highlighted by the superior performance of random forest. The findings help with survival analysis prediction applications in the medical field. Ashfaq et al.[37] used contextual embedding and expert features on actual EHR data, a cost-sensitive LSTM neural network, to reduce unplanned readmissions in congestive heart failure (CHF) patients. The model highlights its practical application by showcasing its better-discriminating ability and possible cost reductions for targeted treatments. Ismael et al.[38] presented a Residual Network-based model that yielded accurate results on a benchmark dataset for classifying brain tumors from MRI images. Khan et al.[39] presented a novel transfer learning-based deep learning system for detecting and classifying breast cancer in cytology pictures, exceeding previous designs in accuracy. Waring et al.[40] examined automated machine learning (AutoML) in the healthcare industry, highlighting the potential, obstacles, and difficulties. Potential healthcare adoption of industrial goods and open-source solutions is examined.

In their systematic literature review (SLR), Mujtaba et al.[41] provided an extensive overview of clinical text categorization over the past five years. The review highlights current research challenges and solutions, offering valuable insights for researchers in this field. Paniagua et al.[42] suggested using deep learning for relation extraction (RE) and named entity recognition (NER) from medical texts, focusing on pharmacovigilance. Their method, which incorporates CNN, CRF, and Bi-LSTM structures, demonstrated cutting-edge functionality in DDI extraction and the eHealth-KD challenge. The authors plan to explore clinical domain embeddings and deeper architectures in future studies. Gargoiulo et al.[43] evaluated various Word Embedding (WE) models on PubMed articles. They explored a deep learning model for extreme text categorization using several classes and labels and a hierarchical label set expansion (HLSE) approach. Their research sheds light on the effectiveness of HLSE and the impact of different WE models. Topaz et al.[44] developed NimbleMiner, an open-source clinical text mining system that utilizes machine learning and NLP to obtain data from health narratives. NimbleMiner outperforms rule-based solutions and is particularly beneficial in fields such as nursing, where NLP advancements are limited. Yadav et al.[45] compared transfer learning, capsule networks, and traditional SVM techniques for CNN-based categorization using tiny datasets of chest X-ray images. Their findings suggest the need for further research on stable transfer learning techniques, evaluation of more powerful CNN models, and incorporation of visualization for clinical applications.

Garg et al.[46] utilized machine learning and NLP to automate the subtyping of ischemic strokes (IS) from electronic health information. The results indicate promising performance, with potential implications for large-scale epidemiologic research. Juhn et al.[47] emphasized the importance of considering unintended effects when utilizing NLP to retrieve information from electronic health records (EHRs). They underline how meticulous design is necessary and the application of health information technology (HIT) in clinical care. Dreisbach et al.[48] focused on using NLP for symptom extraction from electronic patient-authored text (ePAT). They stress the importance of prioritizing patient needs and advocate for in-the-moment symptom evaluation using NLP. Zhang et al.[49] applied text mining and NLP to construction accident records, employing an optimal ensemble model for classification. They recommend future improvements such as data balancing, refining the list of stop words, and exploring more complex models like LSTM neural networks. Kraljevic et al.[50] utilized self-supervised learning to enhance concept extraction and information retrieval from unstructured EHR material using the open-source toolbox MedCAT. Their research demonstrates the successful validation of cross-domain EHR usability across hospitals in London.

The research by Kormilitzin et al.[51] 0.957 was the F1 score attained in electronic health records. They demonstrated the transferability of mental health records from the US intensive care unit to the UK. Model refinement is crucial for trustworthy results in related fields. Locke et al.[52] supported patient outcome prediction, hospital triage, and critical care diagnostic models. Challenges include clinician training and unbiased training data. Integration of clinical practice is planned. Pandey et al.[53] found that deep learning offers enhanced usability for health data with medical imaging and natural language. They explored future directions, challenges, and architecture comparisons. Casey et al.[54] examined using NLP in radiology reports and found significant advancements, especially in deep learning techniques. Gao et al.[55] discussed how to modify BERT for document classification on lengthy clinical texts, pointing out difficulties and recommending that more straightforward models like CNN or HiSAN would work better.

Li et al.[56] suggested a three-phase hybrid approach that combines regular expressions with ABLSTM to categorize medical texts. The technique works better than regular expression-based and ABLSTM models and improves interpretability. Future research aims to expand applications to include more NLP activities. Alhogail et al.[57] suggested using NLP and a graph convolutional network (GCN) to create more accurate results. The model's accuracy rate was relatively high. Subsequent investigations will examine non-English datasets and evaluate their efficacy in mitigating zero-day spear-phishing assaults. Barber et al.[58] assessed if using natural language processing (NLP) to analyze preoperative CT images improves the accuracy of readmission and problem prediction during ovarian cancer surgery. Turchin et al.[59] evaluated the effectiveness of three BERT implementations (regular, BioBERT, and ClinicalBERT) for recognizing intricate medical ideas. BioBERT and ClinicalBERT, trained in the biomedical domain, performed better. Das et al.[60] offered an NLP model based on machine learning to extract neurologic outcomes from clinical notes. It has a high degree of accuracy and can be used to scale neurological research using electronic health record data.

In their study, Han et al.[61] developed a method to automatically extract social determinants of health (SDOH) from clinical records. They improved previous techniques by regularly creating SDOH categories using advanced deep-learning models like BERT. The framework's effective detection of SDOH can significantly enhance healthcare outcomes. Zhou et al.[62] explored using NLP in competent healthcare to analyze human language in various ways. Their article delves into NLP-driven innovative healthcare methodologies, applications, limitations, and prospects. Khanbhai et al.[63] identified issues with care transitions by analyzing 69,285 free-text patient comments from Friends and Family Test (FFT) reports. Recognition of these concerns can guide changes for timely healthcare delivery. Lavanya and Sasikala[64] emphasized the importance of sharing various information, which generates massive unstructured data. They demonstrated that Deep Learning and NLP can improve performance and accuracy in healthcare text categorization. Richard et al.[65] highlighted the significance of detecting protocol deviations early in clinical research. They emphasized that NLP techniques such as TF-IDF, SVM, and word embeddings can aid clinical trial operations by facilitating data-driven decision-making and categorization.

Ubeda et al.[66] utilized NLP-based machine learning models to automate the assignment of medical imaging protocols, which improved accuracy and efficiency. Their work demonstrated high accuracy and practical potential. Linda et al.[67] achieved a high micro-F1 score by effectively categorizing Italian pathology reports using automated natural language processing methods, enhancing efficiency and applicability to a broader range of datasets. Kulshrestha et al.[68] employed clinical notes, NLP, and machine learning to differentiate between the severities of thoracic injuries in trauma situations. Positive outcomes call for further testing in all body areas. Borjali et al.[69] demonstrated that deep-learning NLP models can consistently detect unfavorable occurrences related to hip dislocation in traditional and non-traditional medical narratives, showcasing many applications. Johnson et al.[70] found that NLP demonstrated superior accuracy and sensitivity in recognizing pulmonary embolism episodes compared to ICD-10 codes. Liu et al.[1] provided a convolutional attention network with a medical code prediction application for multi-label document categorization. Among the advances are a deep encoder employing ResSE blocks, concentrated loss for unusual labels, and feature extraction using multi-layer attention. produces cutting-edge outcomes using MIMIC-III and non-English datasets. Subsequent investigations endeavor to evaluate resilience across diverse datasets and enhance attention processes even more. Hu et al.[71] developed the “Squeeze-and-Excitation” (SE) block to make CNNs better by recalibrating channel-wise feature responses. SE blocks significantly improve network performance across datasets, albeit they do come with a little computational cost. Additional complexity and processing load might be limitations. Scalability and applicability to different types of tasks should be studied in future research. Esraa Hassan et al.[72] presented a knowledge distillation model for detecting the picture of Acute Lymphoblastic Leukemia and the contribution of the optimizer as a Nesterov-accelerated adaptive moment estimation optimizer. Abeer Saber et al.[73] proposed an efficient breast tumor detection and classification method using an optimized ensemble model and meta-heuristic algorithms for optimization. Samar Elbedwehy et al.[74] also combined neural networks and advanced optimization approaches to improve the accuracy of kidney disease diagnosis using sensor data, achieving the best performance regarding medical professional-grade diagnostic tasks. Table 1 shows a summary of the findings of the literature. While existing research has shown promising results, more effective and accurate CDC for longer documents is still desired.

References	Methods	Datasets	Advantages	Limitations
2	SVM, RF, MLPNN, and CNN	i2b2 dataset	Improved word embeddings and efficient classification	Not yet evaluated with complex datasets
6	DUAL-CONE X-RAY EXCITATION and IMAGING MODEL	Medical imaging dataset	Better analysis and image classification	Improvements in leveraging other kinds of data in healthcare are needed
22	Synergic deep learning (SDL) model	ImageCLEF-2015, ImageCLEF-2016, ISIC-2016, and ISIC-2017 datasets	This model uses multiple DCNNs along with synergic networks to enable mutual learning	In the future, reinforcement learning algorithms will be needed to automatically search for the number of DCNNs, model parallel computing, and use structural optimization to expand the scale of the SDL model
25	Deep learning	Image datasets	Better segmentation and classification	Novel deep-learning models are to be explored in the future
30	Deep learning and multi-modal medical image segmentation	BraTS dataset	Better feature fusion and accuracy	Feature representation is yet to be improved
38	Deep residual network	Benchmark brain tumor MRI images dataset	Higher level of accuracy	Yet to be improved with larger datasets
41	Supervised machine learning	Free-text clinical report dataset	Better approach in training and testing	Provides valuable insights into clinical data analytics
43	Deep neural network	PubMed dataset	The advantages of DL-based methods include not requiring manual engineering of characteristics and being based on a single DNN architecture, which necessitates a less complex training phase	An intrinsic limitation of word embeddings (WE) is that they cannot explicitly encode grammatical or syntactical information into the resulting word vectors
45	Deep convolutional neural network	Medical image datasets	The best results are achieved through transfer learning of VGG16 with one retrained ConvLayer, slightly surpassing the state-of-the-art result	However, further research is needed due to the limited time. In transfer learning, training a finely tuned deep neural network with unfrozen ConvLayers tends to overfit
55	Deep learning and neural networks	MIMIC-III dataset	Better performance	Unfortunately, these approaches have not been trained on clinical data or released publicly. Further evaluation of these methods will be left for future work
56	Deep learning and neural network	ABLSTM model	The proposed approach combines the interpretability of rule-based algorithms with the computational power of deep learning for a production-ready scenario	Future investigations will use ABLSTM and regular expression rules for other NLP tasks, such as entity recognition and relation classification, in various domains to induce general patterns of named entities
65	NLP and SVM	PD dataset	Improved performance in data analytics	One of the main challenges of our current work is the intensive resources and expertise required to label PDs in the training set
66	ML models	CT and MRI dataset	Improvement in medical image analytics	Additional pre-processing steps are needed, such as negation detection, part-of-speech processing, and clinical ontology assignment
1	EffectiveCAN	MIMIC-III, Dutch and French datasets	Improved efficiency in clinical document classification	CAN model could be further enhanced

Table 1. Summary of literature findings.

Materials and methods

This section describes the underlying mechanics and the suggested methods for automatically categorizing medical records. Multi-label clinical document (MCD) classification is a job in natural language processing for which the approach is intended. A machine learning model is taught to classify clinical documents into many labels or categories in MCD classification. These labels can represent different medical conditions, treatments, or other relevant information in the document. The objective is to accurately classify the document into one or more categories based on its content, enabling efficient organization and retrieval of medical information.

Materials

This study utilized four publicly available datasets[1]: English, Dutch, MIMIC-III-full, and MIMIC-III-50. These datasets are designed for clinical document classification tasks, where each instance represents a clinical discharge summary—a narrative document summarizing a patient's hospital stay, diagnosis, procedures, and follow-up instructions. These summaries are ideal for multi-label classification due to their rich medical content and diverse diagnostic annotations.

MIMIC-III (Medical Information Mart for Intensive Care III) is a de-identified health dataset comprising approximately 40,000 ICU patients admitted to the Beth Israel Deaconess Medical Center between 2001 and 2012. It includes diverse data types such as demographics, lab tests, medications, imaging reports, clinical notes, and discharge summaries. For this study, only discharge summaries were used.

MIMIC-III-full contains 47,724 training instances, 1632 validation instances, and 3372 test instances. The average document length is approximately 1485 words (not a fixed length), and each document is associated with multiple diagnostic labels. There are 8922 distinct labels—each corresponding to ICD-9 diagnosis or procedure codes—with an average of 15.9 per instance. These labels represent a wide variety of medical conditions and procedures. Given this high dimensionality, class imbalance is a notable concern, as a small set of labels appears frequently while many others are rare. This imbalance was addressed using focal loss, emphasizing learning from underrepresented classes.

MIMIC-III-50 is a subset of MIMIC-III. It focuses on the 50 most frequent ICD-9 codes, resulting in a more balanced label distribution. It includes 8067 training, 1574 validation, and 1730 test instances, with an average of 1530 words and 5.7 labels per instance.

The Dutch dataset comprises 2511 training, 313 validation, and 313 test instances, each with an average of 4958 words and five labels drawn from 144 total ICD-related codes. The French dataset (not previously mentioned by name) contains 22,540 training, 2836 validation, and 2827 test instances, with 940 total labels and an average of 11 labels per document.

All datasets were split randomly at the document level while ensuring that documents from the same patient did not appear in multiple splits, thus preventing data leakage. No time-based or admission-based constraints were applied. Text preprocessing steps include Lowercasing, Tokenization, Removal of non-alphanumeric characters and extra white spaces, Stopword removal, and Punctuation stripping. Stemming or lemmatization was not applied to preserve medical terminology. Documents were truncated or padded to a maximum sequence length (e.g., 2000 tokens) to ensure consistent input dimensions.

The study focused solely on unstructured clinical text (discharge summaries) and did not incorporate structured metadata features such as patient demographics or lab results. Additionally, no explicit feature selection or dimensionality reduction was applied before model training. Instead, the deep learning model inherently learned high-level features through its convolutional and attention-based architecture. Figure 1 shows data distribution details.

The datasets used in this study, particularly MIMIC-III and its variants, are comprehensive databases originated in Beth Israel Deaconess Medical Center containing clinical discharge summaries with great detail. These abstracts are multi-disease rich and widely utilized for clinical document classification tasks. To reinforce knowledge and educational value, a template link of a discharge summary with its relevant ICD-9 codes will be added as a supplement or as in Table 2.

The label space is high-dimensional; for instance, the MIMIC-III-full dataset has 8922 unique ICD-9 codes. The labels are heavily imbalanced, with only a few codes appearing often (frequency) and the rest rare. To bypass this, we used focal loss, which down-weighs easy-to-classify (i.e., under-represented) labels by tuning it towards hard-to-classify examples. This reduces performance drop and infrequent classes.

We adopted a random split approach, ensuring that no discharge summaries for the same patient are present in multiple splits to prevent data leakage. To maintain consistency and enable comparison with previous clinical NLP work, we used the standard split ratio (approximately 80:10:10) for all datasets. This is to capitalize on the systematic approach.

Methods

Clinical documents can vary widely in content and structure. They may include medical reports, patient notes, discharge summaries, imaging reports, pathology reports, etc. Each document may contain information about medical conditions, treatments, and symptoms. Unlike single-label classification, where each document belongs to a single category, clinical documents often contain information pertinent to multiple medical conditions, procedures, or symptoms. There can be complex relationships between different labels. For instance, certain medical conditions might co-occur frequently, or specific symptoms might indicate multiple conditions. Clinical documents can be lengthy, sparse, and noisy, challenging feature extraction and classification. Moreover, documents may vary in writing styles, terminology, and level of detail. Multi-label clinical document classification is a task where machine learning models are trained to assign multiple labels or categories to clinical documents based on their content. This task is essential in healthcare and medical research for efficiently organizing, indexing, and retrieving clinical information. We propose a methodology for classifying clinical

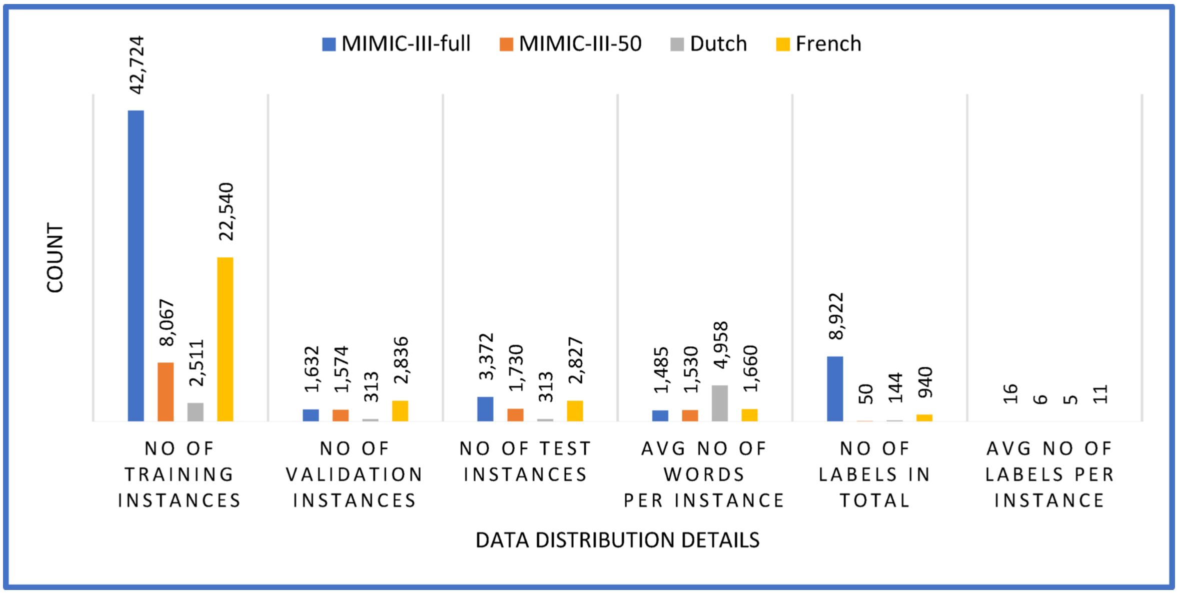


Fig. 1. Datasets and data distribution dynamics.

Discharge summary excerpt	ICD-9 code(s)	Description
"The patient is a 67-year-old male admitted with shortness of breath and chest discomfort. ECG showed ST elevation. Coronary angiography revealed triple vessel disease. The patient underwent CABG successfully. Diabetes and hypertension managed during the stay."	410.90250.00401.9V45.81	Acute Myocardial Infarction Type II Diabetes Mellitus Essential Hypertension Status post-CABG

Table 2. Sample clinical document and ICD-9 code labels.

documents automatically and efficiently. The proposed deep learning model, Enhanced Effective Convolutional Attention Network (EECAN) with the squeeze-and-excitation inception module, is designed for multi-label clinical document classification. EECAN is an extension of the EffectiveCAN model proposed in[1]. Figure 2 depicts the general layout of the proposed approach.

Classifying clinical documents starts with the Clinical Document Dataset, which undergoes Data Transformation. The transformed data then undergoes Encoding. After encoding, an Attention mechanism is applied to highlight important characteristics of the information. Finally, the refined Information is entered into a Classification model to generate the Clinical Document Classification Results. This process aims to efficiently transform, encode, and classify clinical documents to achieve accurate results.

Enhanced effective convolutional attention network

Figure 3 illustrates the architecture of the proposed Enhanced Effective Convolutional Attention Network (EECAN), influenced by EffectiveCAN[1]. Deep convolution-based encoders are utilized, with an input layer processing raw document strings to learn meaningful representations of the document texts. Following this, an output layer generates the final predictions. Additionally, an attention component selects the most significant textual elements, which are then used to create label-specific representations for each label.

The model's principal objective structure is to improve predictions on multi-label classification tasks, and it achieves this in three ways: (1) by generating meaningful representations for input texts; (2) by selecting relevant information for label prediction from these text representations; and (3) by preventing overconfidence on frequently appearing labels. To accomplish this, we first introduce a squeeze-and-excitation (SE)—Inception module into the convolution-based encoder, which helps in producing high-quality representations of the document content. The encoder comprises several encoding blocks designed to capture text patterns of varying lengths and increase the receptive field. Next, instead of simply selecting the encoder layer result, we use attention to extract all the encoding layer outputs and choose the attributes for each label that are the most informative. Finally, we include the binary cross-entropy loss and the focal loss to handle the long-tail distribution of labels and ensure the model works well on common and uncommon labels.

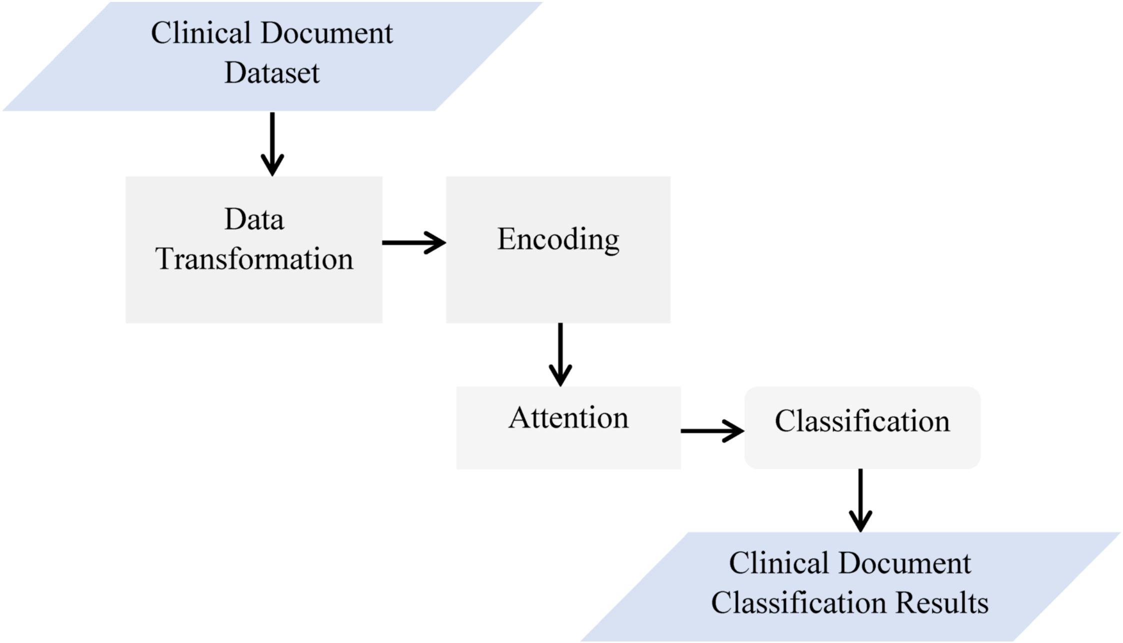

Clinical Document Dataset
Data Transformation
Encoding
Attention
Classification
Clinical Document Classification Results

Fig. 2. Outline of the proposed approach for clinical document classification.

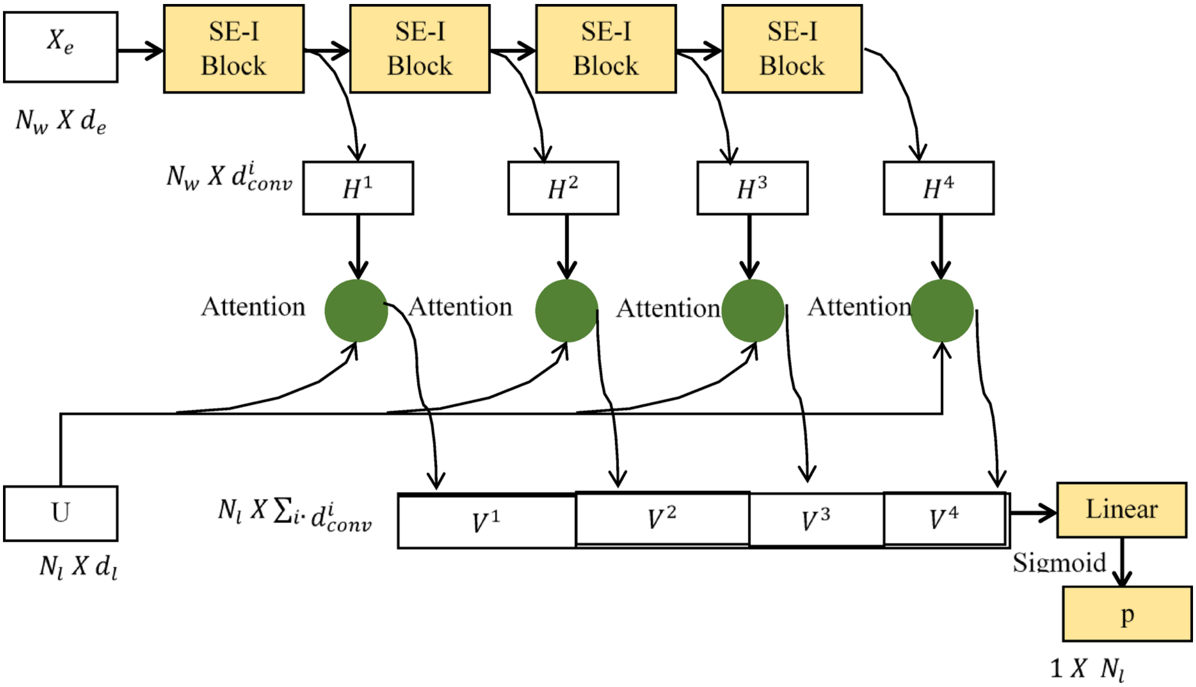


Fig. 3. The architectural overview of the proposed Enhanced Effective Convolutional Attention Network (EECAN).

Input layer

Each word in a string of words is mapped to a size. d_e word embedding by our model. A word embedding matrix will be the input, supposing that the document contains N_w words.

$$X_e = [x_1, \ldots, x_{N_w}] \in \mathbb{R}^{N_w \times d_e}.$$

We upgraded the EffectiveCAN model by substituting the Res-SE block with the SE-I block to improve performance in document classification. The work influenced this change in[71]. Figure 4 illustrates the architecture

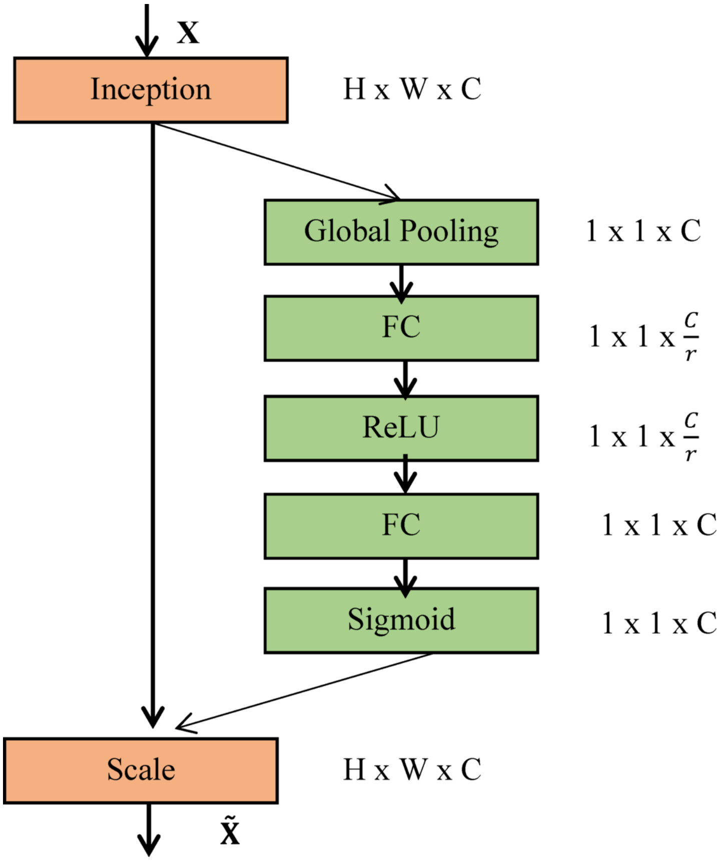


Fig. 4. The structure of the Squeeze-and-Excitation Inception (SE-I) block.

of a Squeeze-and-Excitation Inception (SE-I) block. It begins with an Inception module that takes an input of size H × W × C (height, width, channels). The output is then globally pooled to a 1 × 1 × C tensor, two ultimately linked layers came next with ReLU and sigmoid activations. The output of the sigmoid layer is then used to scale the original input, giving rise to an output equal in size to the input. This block models the interdependencies across the channels to dynamically adjust channel-wise feature responses. This research provides recommendations on improving the credibility of the findings by integrating the Squeeze-and-Excitation Inception (SE-I) module to capture more rigorous shape representations dynamically. This recalibration through squeeze and excitation is a two-step process. The squeeze operation is designed to gather information from global spatial information by applying global average pooling, resulting in a single descriptor of the feature maps along the channel dimension expressing the input data's international context. After that, this descriptor is fed to the excitation operation, composed of two fully connected layers with non-linear activation functions, which learn the relevance of each channel. This set of operations with the output is used to scale the original feature map, amplifying the most representative channels and suppressing the irrelevant ones. The SE-I module substantially enhances the model's capacity to characterize input data, allowing it to concentrate on crucial features and ignore noise and irrelevant data. This type of selective attention helps in multi-label clinical document classification, given that clinical documents usually contain overlapping and co-existing clinical concepts. This allows the model to learn discriminative and context-aware features better by assigning different importance to different channels, thus improving classification performance. Moreover, the seamless interaction of the SE-I module effectively solves the difficulty of dealing with long and intricate clinical documents by enhancing the model's sensitivity to important subtle features. Recalibration helps learn label-specific representations, thus improving the classification of instances with many highly interconnected labels. In addition, the SE-I module is also computationally efficient, adding a negligible overhead while providing substantial performance gain.

Convolutional encoder

We utilize a convolution-based encoder comprising multiple SE-I blocks to process the input word embeddings Xe, which are then transformed into meaningful content representations. Figure 4 depicts how each SE-I block consists of different layers. Recently, integrating self-attention modules into transformer-based models has proven effective in text categorization tasks[75–77]. Therefore, unlike self-attentional encoders, we opt for convolutional encoders in our applications for two reasons: (1) Input sentences are often linked to ICD code predictions, and clinical papers are typically lengthy (the average MIMICIII document is 1500 words long). Convolutional techniques efficiently combine word spans' content and offer insightful insights for subsequent predictions. For simulating lengthy texts, a convolutional encoder outperforms a self-attention encoder in terms of efficiency in both time and space.

Squeeze-and-excitation inception block

The architectural overview found in Fig. 4 is the SE-I block diagram. A transformation F_{tr} translating an input $X \in \mathbb{R}^{H' \times W' \times C'}$ to feature mappings $U \in \mathbb{R}^{H \times W \times C}$ may be used to construct an excitation-squeeze block or a CPU. The following notation treats F_{tr} as a convolutional multiplier, and we use V $= [v_1, v_2, \ldots, v_C]$ to represent the learned collection of filter kernels, where v_C represents the c-th filter's parameters. After that, we may express the resultant as U $= [u_1, u_2, \ldots, u_C]$, where

$$u_C = v_C * X = \sum_{s=1}^{C'} v_c^s * X^s. \tag{1}$$

Here * denotes convolution, $v_C = \left[v_c^1, v_c^2, \ldots, v_c^{C'}\right]$, X = [$X^1, X^2, \ldots, X^{C'}$] and $u_C \in \mathbb{R}^{H \times W}$.

One channel of v_C is represented by a 2D spatial kernel named v_c^s, which operates on the relevant channel of X. Bias words are removed to simplify the notation. Though implicitly contained in v_C channel dependencies are mixed up given that all channels are added together to yield the result; the filters may be combined using the local spatial correlation collected except for the topmost layers, convolution models implicit, and local channel interactions by nature. To increase the network's sensitivity to incoming features that later transformations may use, we anticipate explicitly modeling channel interdependencies will improve the learning of convolutional features. To prepare the data for the next transformation phase, we must adjust the filter responses in two stages—squeeze and excitation. This will require access to global data. Initially, we must analyze the signal for each channel to address any channel dependency issues in the output characteristics. Each learning filter in the transformation output U has a limited operational range. No unit can utilize contextual information beyond its local receptive fields. We suggest condensing global geographical information with a channel description to mitigate this problem. Global average pooling is used to generate channel-wise data to achieve this. Formally, decreasing U via its dimensions in space W × H yields a statistic z∈$\mathbb{R}^C$. Accordingly, one may locate the c-th component of z from:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_C(i, j). \tag{2}$$

To achieve the desired transformation, we believe U is a set of local descriptors with data that fairly depict the whole picture. This approach has been commonly used in previous feature engineering work[78,79], and[80]. While more complex methods could be considered, we opt for the simple global average pooling aggregation technique. To capture channel dependencies effectively, we perform an additional step after the squeeze operation to utilize the information gathered during the initial step. To achieve this, the function must meet two requirements: it needs to be adaptable, specifically able to learn a non-linear relationship between the channels, and it must exhibit interpersonal skills to ensure that different channels can be highlighted without requiring a one-hot activation. We use a primary gating mechanism with a sigmoid activation to meet these requirements.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \tag{3}$$

$W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, where δ denotes the ReLU[81] function. By creating a bottleneck, our gating mechanism is parameterized by two fully connected (FC) layers that surround the non-linearity: a dimensionality-reduction layer that returns to the channel dimension of the transformation output U. This dimensionality-increasing layer has a reduction ratio of r and a ReLU. This constrains model complexity and aids in generalization. Rescaling U using the activations yields the block's final output:

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c u_c, \tag{4}$$

where $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_C]$ and $F_{scale}(u_c, s_c)$ means to multiply the scalar multiplex by the channel s_c and the feature map $u_C \in \mathbb{R}^{H \times W}$. Channel weights are mapped to the input-specific descriptor z via the excitation operator. As such, SE blocks add dynamics conditioned by nature on the input. This may be considered a self-awareness feature on networks that connections extend beyond the convolutional filters' local receptive response area.

The SE Inception module differentiates itself by modulating the feature representation by applying the SE Inception module that adaptively re-weights the channel-wise feature responses, allowing the model to emphasize the more pertinent features and down-weigh the ornaments. The module combines the global pooling and excitation operations, which simultaneously seize local and global dependencies, to improve the discriminative ability of the feature maps. Such a new calibration allows the model to better extract label-dependent characteristics, especially in multi-label classification tasks where relations between the labels must be plotted. As a result, the SE Inception module dramatically enhances the classification accuracy for complex, long-tail label distributions by refining the feature representation and, in turn, improving the robustness of the entire model.

Attention

We use label-wise attention to create label-specific representations from H[82]. We undertake multi-layer attention, which pays attention to the outputs of each SE-I block in our convolutional encoder as it is composed of several SE-I blocks that provide the document texts in many size formats. The encoder extracts a rich feature space from which each label may select the most relevant attributes. Assuming that the label embedding matrix is represented by $U \in \mathbb{R}^{N_l \times d_l}$, where d_l is each label's embedding size and N_l is the sum of all the labels. Responding to the output of the ith SE-I layer $H^i \in \mathbb{R}^{N_w \times d^i_{conv}}$ and prevent dimension mismatch, U is initially assigned to $U' \in \mathbb{R}^{N_l \times d^i_{conv}}$ via a filter-size-1 convolutional layer. Next, those who decide the attention weights are expressed in Eq. (5).

$$A^i = softmax\left(U' \cdot H^{i^T}\right) \tag{5}$$

In this instance, the weight vectors indicate the informativeness of the text representations in H for the *j*th label are found in each of the *j*th columns of $A^i \in \mathbb{R}^{N_l \times N_w}$. We then create the representations that are unique to each label: The formula $V^i = A^i \cdot H^i$ denotes that the *j*th column in $V^i \in \mathbb{R}^{N_l \times d^i_{conv}}$ represents the label-specific representation of the *j*th label, which is produced by the *i*th SE-I layer output.

After concatenating the depictions particular to each label, we repeat the attention procedure for each output of the SE-I layer.

$$V = concat\left(V^1, \ldots, V^{N_{SE-I}}\right) \tag{6}$$

N_{SEI} represents the quantity of SE-I blocks. The final prediction will be based on the resultant $V \in \mathbb{R}^{N_l \times \sum_i d^i_{conv}}$.

It can be challenging to train a multi-layer attention model, particularly for deep networks, when there is abundant label space in the application region but insufficient data. Consequently, we also test sum-pooling attention, where each convolutional layer is first transformed to match the final layer's dimension. After that, all the layers are combined, and the result is examined. The final prediction is made using the resultant expression $V' \in \mathbb{R}^{N_l \times d^{last-layer}_{conv}}$.

Output layer

We first use a fully connected layer to acquire the label-specific representations. We sum-pool the data and use it to find the probability for every label using a sigmoid transformation:

$$p = sigmoid\left(pooling\left(W_{fc} \cdot V^T + b_{fc}\right)\right) \tag{7}$$

where $b_{fc} \in \mathbb{R}^{N_l}$ and $W_{fc} \in \mathbb{R}^{\sum_i d^i_{conv} \times N_l}$. Given the document contents, the *j*th number in p represents the expected likelihood that the *j*th label would be present.

Loss function dynamics

Binary cross-entropy loss is a popular choice for the loss function in MLDC model training. Assuming that p represents the expected probability, and y represents using the binary cross entropy loss as the ground truth label,

$$L_{BCE}\left(p_t\right) = -logp_t.$$

We also use the focused loss to address the labels' long-tail distribution. This method dynamically reduces the loss associated with correctly classified labels by adding a binary cross entropy loss standard with a weight term. The primary loss is

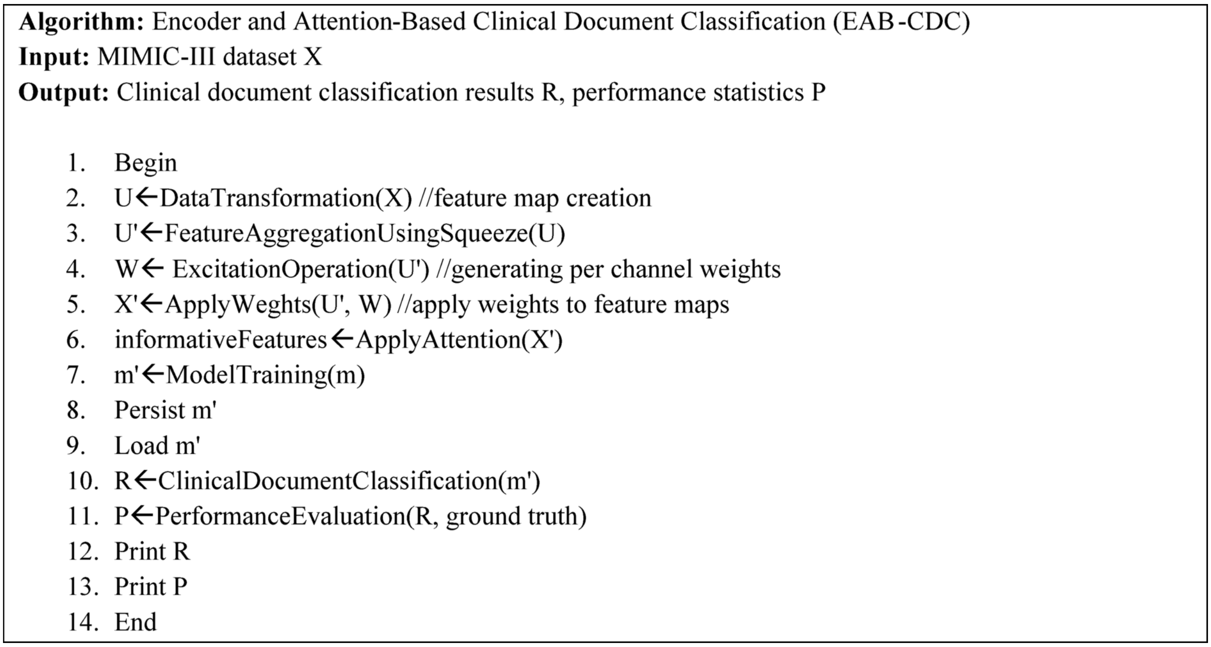
$$L_{FL}\left(p_t\right) = -\left(1 - p_t\right)^{\gamma} logp_t \tag{8}$$

In this case, adjusting the γ parameter may alter the down-weighting strength. When p_t is large, the weight term $\left(1 - p_t\right)^{\gamma}$ suppresses the loss from well-classified labels, which causes the model to be biased toward labels with inaccurate predictions. Utilizing attention loss right from the start of training is not good since it prioritizes correcting misclassified uncommon labels at the expense of frequent labels. Instead, it's common to start training models with binary cross-entropy loss to help them learn large-scale features and achieve high performance on frequent labels. Once the model's performance plateaus, we can further introduce focus loss to improve predictions for atypical labels.

To cater to the label long-tail distribution, the model uses focal loss to reduce the loss contribution of easy-to-class examples while focusing more on hard-to-class examples. This makes the model more sensitive to the minority classes and avoids the skew from frequent labels. Moreover, performance metrics, such as macro-averaged F1-score, were reported to ensure the abundance of a small number of classes did not dominate performance metrics.

Proposed algorithm

We have developed an Encoder and Attention-Based Clinical Document Classification (EAB-CDC) algorithm for accurately categorizing clinical documents using the MIMIC-III dataset. This algorithm is based on the proposed EECAN model. It utilizes a squeeze operation for feature aggregation, an excitation operation for generating per-channel weights, and an attention mechanism to focus on the most informative features. Our goal with this method is to enhance the categorization of clinical documents by highlighting the most relevant parts of the data, ultimately leading to better performance in clinical document classification tasks.



Algorithm: Encoder and Attention-Based Clinical Document Classification (EAB -CDC)

Input: MIMIC-III dataset X

Output: Clinical document classification results R, performance statistics P

```
Begin
U←DataTransformation(X) //feature map creation
U'←FeatureAggregationUsingSqueeze(U)
W← ExcitationOperation(U') //generating per channel weights
X'←ApplyWeghts(U', W) //apply weights to feature maps
informativeFeatures←ApplyAttention(X')
m'←ModelTraining(m)
Persist m'
Load m'
R←ClinicalDocumentClassification(m')
P←PerformanceEvaluation(R, ground truth)
Print R
Print P
End
```

Algorithm 1. Encoder and attention-based clinical document classification (EAB-CDC).

Algorithm 1 is designed to process the MIMIC-III dataset (X) and generate performance statistics (P) and classification results (R). The algorithm begins with a data transformation step to create maps with features from the given data collection. Following an excitation procedure to produce per-channel weights, these feature maps are further subjected to feature aggregation using a squeeze operation. Next, the feature maps are subjected to the weights. The algorithm then highlights the most pertinent portions of the data by paying close attention to the most valuable attributes. A model training phase is then carried out, and the trained model is stored for use at a later time. Once the model is loaded, it is used to classify clinical documents and evaluate their performance against the ground truth data. The classification results and performance statistics are then printed. The EAB-CDC algorithm is a structured approach to classifying clinical documents using a combination of feature transformation, aggregation, and attention mechanisms to improve document classification accuracy. The algorithm's effectiveness is validated through performance evaluation, and the results are presented clearly and concisely.

Evaluation methodology

The primary goal of computer-aided clinical coding is to minimize human involvement by enabling models to automatically assign the correct codes from the entire label space rather than merely suggesting a ranked list or the top-N probable codes. Many studies evaluate performance based on the top 50 most frequent ICD codes, which is insufficient in real-world clinical scenarios where patients often have complex, multi-condition profiles. In MIMIC-III, for example, only about one-third of the actual codes in each patient encounter are captured within the top 50 codes, highlighting the limitations of such a restricted evaluation scope.

Furthermore, there is significant variation in the number of labels per clinical document—ranging from 1 to 79 codes per instance in MIMIC-III. Over 43% of encounters have more than 15 codes. Metrics like Precision@K (P@K), Recall@K (R@K), and Rank Precision@K (RP@K) are unreliable in such settings because they assume a fixed number of relevant labels (K) per document. This fixed-K assumption breaks down in clinical text, where label counts vary widely, leading to skewed or inflated performance metrics, especially for documents with very few or very many labels. As noted by[83,84], limiting evaluation to K labels can distort recall and introduce bias, particularly in documents with sparse label sets. We adopt micro- and macro-averaged precision, recall, and F1-score standards for multi-label classification tasks to address this. Micro-averaged metrics aggregate all true/false predictions across all labels and instances before computing the scores, favoring frequent codes and reflecting the model's global performance across the entire dataset. Macro-averaged metrics compute scores independently for each label and then average them, placing equal weight on rare and frequent codes. This is valuable for understanding model behavior on underrepresented classes, though its use as a sole metric may distort the perception of overall effectiveness.

Given the skewed label distribution and the long-tailed nature of clinical codes, Micro F1 is critical, as it emphasizes accurate overall prediction across all labels and is better aligned with real-world coding productivity and system deployment requirements. Although previous research often underreported precision and recall on MIMIC-III, we include them here to offer a clearer picture of both sensitivity (recall) and specificity (precision) in the model's performance—critical for downstream clinical use, where both overcoming and under coding carry risks.

Finally, while our main results focus on micro-level metrics, we also report macro F1, precision, and recall across the MIMIC-III and non-English datasets to support comprehensive comparison with prior work. Interpretability of model errors—such as identifying whether a mistake is a false positive or a false negative—is also essential for human-in-the-loop systems. While not the focus of this paper, future work will include detailed error-type analyses to support transparency and assist human coders during model validation and deployment.

Precision, Recall, and F1-score were chosen due to their appropriateness for multi-label classification tasks and capacity to represent performance on imbalanced datasets. The F1 score precisely balances the trade-off between precision and recall, so it is a more informative measure than accuracy, especially when there is a class imbalance. These are commonly used measures in clinical NLP benchmarks and are more realistic evaluations of model effectiveness across all classes.

To enhance our evaluation, we acknowledge that no single metric is sufficient when it comes to clinical document classification[25]. Thus, alongside Micro F1—which captures overall performance well and deals with label imbalance—we also report Macro F1 to emphasize behavior on rare codes. It should be noted that hierarchical F1 and coverage errors are not incorporated in this study. These are suitable metrics for ICD coding tasks to enable comparison between partial correctness (hierarchical F1) and ranking quality (coverage error). Additionally, precision and recall, the underreporting of which in studies using MIMIC-III given their importance for interpretability, which can also be used to perform, among others, the types of error analysis that is relevant when an automated coding system is helping or augmenting human coders. Further development of the study approach will include hierarchical metrics and detailed error analysis to conform to real-world clinical coding system standards and industry benchmarks.

Experimental results

Using the data sets from section "Materials", the section displays the outcomes of the experiments. Except for keeping numerical values between one and ten since they are essential for coding, we adhere to the work in[82] for preprocessing methodology. Using the preprocessed texts for MIMIC-III and the non-English sets, the word2vec CBOW technique pretrain the word embeddings of size d_e = 100 and 200, respectively. While the non-English sets were given two different sequence lengths—2500 and 3500—all MIMIC papers are trimmed to a maximum size of $w_{max} = 3500$. We utilized the Ray Tune library to determine the ideal hyperparameter values[86]. Following the input embedding layer, each SE module's dropout probability (q), power term (γ) in the focal loss function, out-channel size (d_conv) in the convolutional layer, and filter size (k) were adjusted. To narrow down the search space, we establish $d^1_{conv} = d^2_{conv}$, $d^3_{conv} = d^4_{conv}$, and $k^1 = k^2$, $k^3 = k^4$. Table 3 compiles their ideal values for various trials. Our studies employ The Adam optimizer with a 0.00015 starting learning rate and four SE-I blocks. The effectiveness of the suggested model, EECAN, is evaluated against several other models for clinical document classification. The newest models available include CAML[82], C-LSTM[85], MSATT-KG[72], MultiResCNN[73], HyperCore[74], and LAAT[87]

Document classification in the setting of unstructured clinical text, specifically discharge summaries, is the sole focus of this study. The current approach did not include structured metadata—like patient demographics, lab results, or vital signs. Future works can investigate using structured and unstructured data for classification boosting performances, contextual understanding, and interpretability of models for deployment in real-life healthcare applications.

Using the MIMIC-III full dataset, Table 3 displays the outcomes of clinical document categorization models that have been suggested and those that are currently in use.

Model	AUC	F1	P@k 8	P@k 15
CAML	0.986	0.539	0.709	0.561
DR-CAML	0.985	0.529	0.690	0.548
MSATT-KG	0.992	0.553	0.728	0.581
MultiResCNN	0.986	0.552	0.734	0.584
HyperCore	0.989	0.551	0.722	0.579
LAAT	0.988	0.575	0.738	0.591
JointLAAT	0.988	0.575	0.735	0.590
EffectiveCAN (Multi-layer attention)	0.989	0.581	0.604	0.755
EffectiveCAN (Sum-pooling attention)	0.988	0.589	0.758	0.606
EECAN (Multi-layer attention)	0.998	0.585	0.609	0.761
EECAN (Sum-pooling attention)	0.997	0.592	0.765	0.615

Table 3. Experimental results, using the dataset MIMIC-III-full, of clinical document classification.

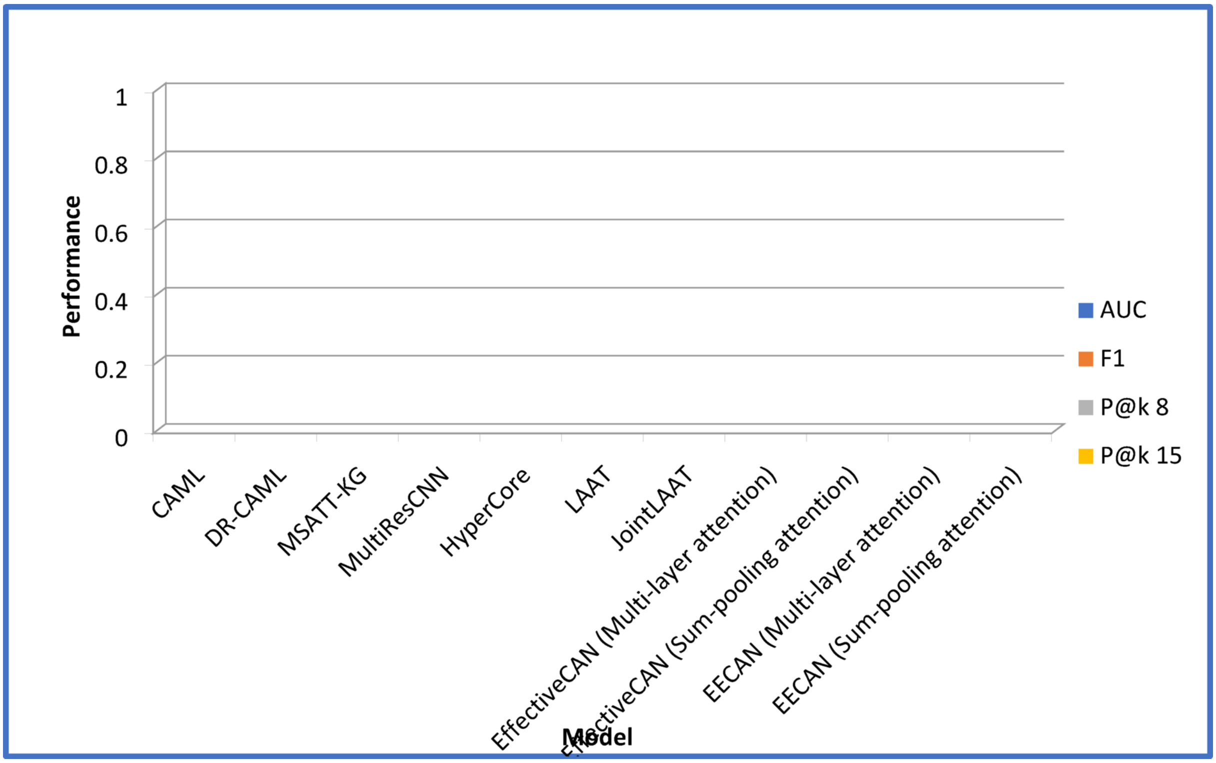


Fig. 5. Results of clinical document classification exhibited by various models using the MIMIC-III-full dataset.

Model	AUC	F1	P@k 5
C-LSTM-Att	0.900	0.532	
CAML	0.909	0.614	0.609
DR-CAML	0.916	0.633	0.618
MSATT-KG	0.936	0.684	0.644
MultiResCNN	0.928	0.670	0.641
HyperCore	0.929	0.663	0.632
LAAT	0.946	0.715	0.675
JointLAAT	0.946	0.716	0.671
EffectiveCAN (Multi-layer attention)	0.945	0.717	0.664
EffectiveCAN (Sum-pooling attention)	0.938	0.702	0.656
EECAN (Multi-layer attention)	0.954	0.725	0.675
EECAN (Sum-pooling attention)	0.942	0.712	0.678

Table 4. Experimental results of MIMIC-III-50 dataset.

Figure 5 compares the effectiveness of many models as measured by four metrics: AUC (Area Under the Curve), F1 score, P@k 8 (Precision at top 8), and P@k 15 (Precision at top 15). Each model is shown on the x-axis, while performance metrics are on the y-axis. All models exhibit high AUC values, consistently above 0.8, and some are close to 1.0, indicating overall solid performance. About 0.5 to 0.7 is the range of F1 scores, indicating a somewhat successful balance between recall and accuracy. The precision metrics P@k 8 and P@k 15 vary across models but generally range from 0.6 to 0.8, indicating how well the models rank relevant items at these specific cutoffs. The models compared include CAML, DR-CAML, MSATT-KG, MultiResCNN, HyperCore, LAAT, JointLAAT, and variations of EffectiveCAN and EECAN. This detailed comparison highlights the advantages and disadvantages of every model in terms of specific performance indicators.

Table 4 presents the results of clinical document classification models, both existing and proposed, using the MIMIC-III-50 dataset.

AUC (Area Under the Curve), F1 score, and P@k 5 (Precision at top 5) are the three measures we use in Fig. 6 to compare the performance of various models. Performance measurements are displayed on the y-axis, while each model is represented on the x-axis. All models demonstrate high AUC values, consistently above 0.8,

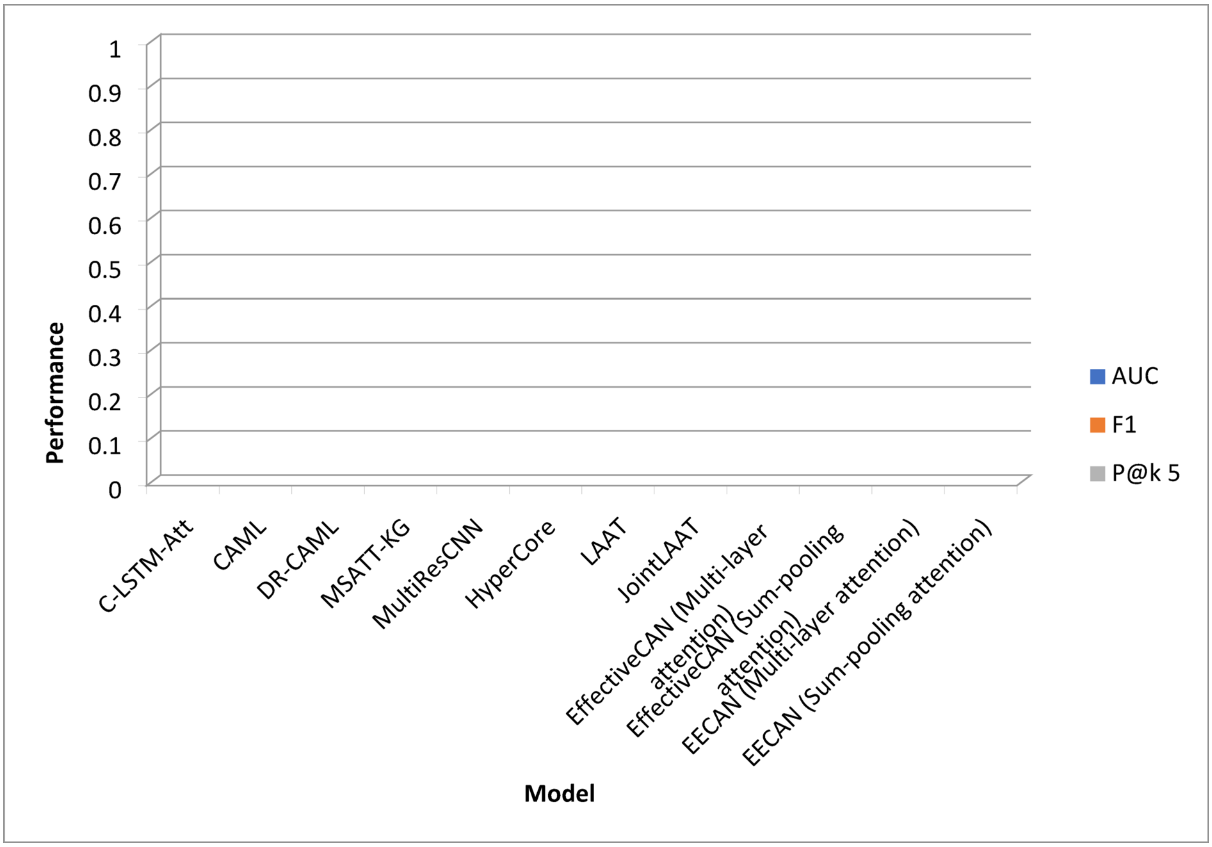


Fig. 6. Results of clinical document classification exhibited by various models using the MIMIC-III-50 dataset.

Model	Dutch			
	# Labels	Precision	Recall	F1
XLM-RoBERTa	50	0.725	0.289	0.413
MultiResCNN	50	0.458	0.639	0.534
EffectiveCAN	50	0.822	0.760	0.790
EffectiveCAN (3500)	50	0.873	0.777	0.822
EffectiveCAN (3500)	144	0.844	0.732	0.784
EECAN (Multi-layer attention)	55	0.882	0.786	0.836
EECAN (Sum-pooling attention)	150	0.852	0.745	0.795

Table 5. Experimental results of the Dutch dataset.

with some approaching nearly 1.0, indicating overall solid performance. F1 scores range between 0.4 and 0.6, suggesting moderate effectiveness in balancing precision and recall. Precision at top 5 (P@k 5) metrics display variation across models, generally falling between 0.5 and 0.7, reflecting how well the models rank relevant items at this specific cutoff. The compared models include C-LSTM-Att, CAML, DR-CAML, MSATT-KG, MultiResCNN, HyperCore, LAAT, JointLAAT, EffectiveCAN (in both Multi-layer and Sum-pooling attention variants), and EECAN (in both Multi-layer and Sum-pooling attention variants). This detailed comparison highlights the advantages and disadvantages of every model using various performance measures.

Table 5 presents the results of clinical document classification models, both existing and proposed, using the Dutch dataset.

In Fig. 7, we compare the performance of several models across three metrics: Accuracy, Memory, and F1 rating. The y-axis displays the performance measures, while the x-axis displays each model. The models we included are XLM-RoBERTa, MultiResCNN, EffectiveCAN, EffectiveCAN (trained on 3500 samples), and EECAN (in both Multi-layer and Sum-pooling attention variants). Notably, the EffectiveCAN models, especially those trained on 3500 samples, demonstrate high performance across all metrics, with Precision and Recall values generally above 0.8 and high F1 scores. MultiResCNN displays moderate performance, with Precision and Recall around 0.5 to 0.6 and a slightly lower F1 score. XLM-RoBERTa, while having the highest precision among the models, shows lower recall and F1 scores, indicating a disparity between the relevance of retrieved

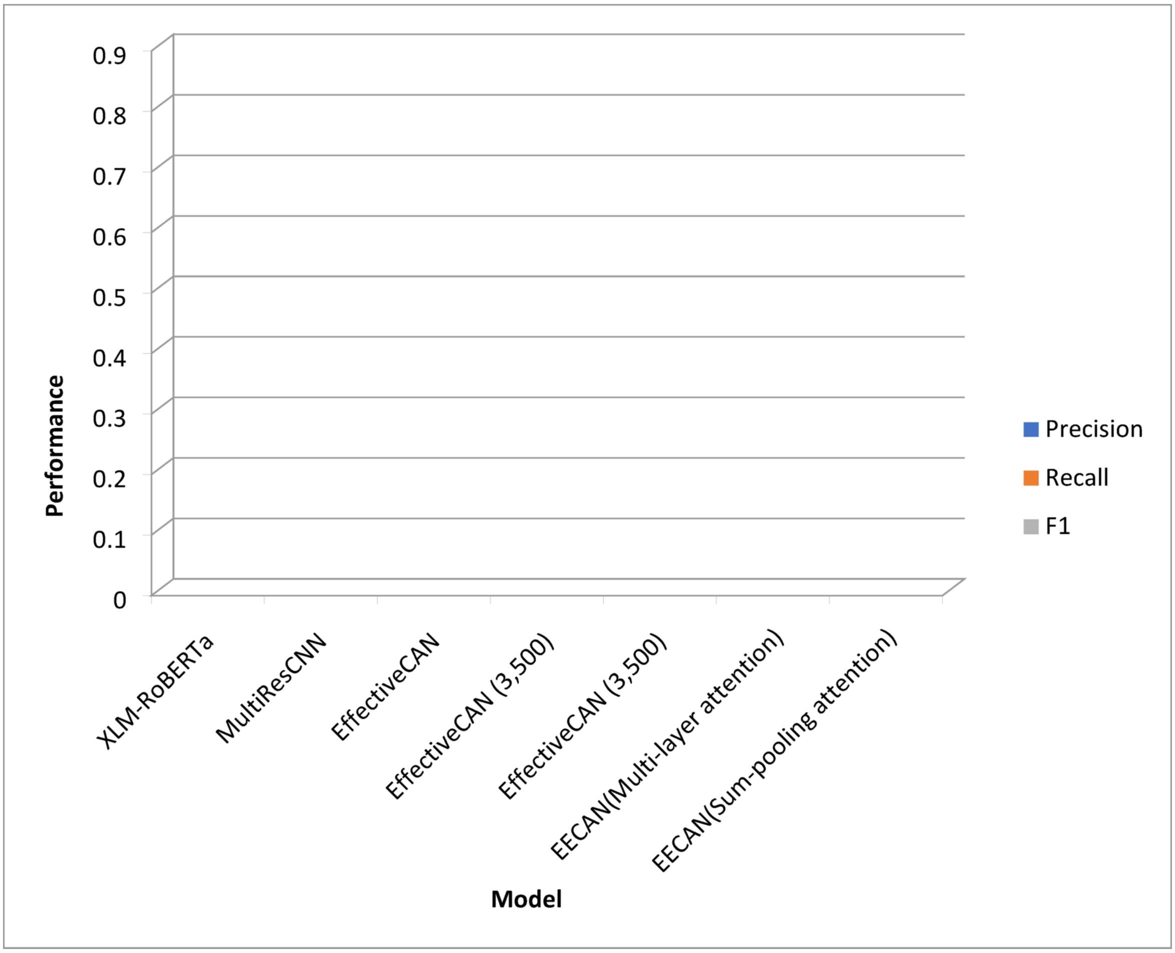


Fig. 7. Results of clinical document classification exhibited by various models using the Dutch dataset.

	French			
Model	**# Labels**	**Precision**	**Recall**	**F1**
XLM-RoBERTa	50	0.606	0.426	0.500
MultiResCNN	50	0.631	0.607	0.619
EffectiveCAN	50	0.692	0.620	0.654
EffectiveCAN (3500)	50	0.705	0.636	0.669
EffectiveCAN (3500)	940	0.583	0.493	0.534
EECAN (Multi-layer attention)	55	0.716	0.646	0.675
EECAN (Sum-pooling attention)	950	0.596	0.499	0.548

Table 6. Experimental results of the French dataset.

instances and the total relevant instances. This detailed comparison highlights the specific strengths of each model, highlighting the satisfactory performance of the accuracy, recall, and F1 score of Enhanced CAN models across different metrics.

Table 6 presents the results of clinical document classification models, both existing and proposed, using the French dataset.

Three metrics—Preciseness, Recall, and F1 score—are used in Fig. 8 to compare the performance of several machine learning models. The evaluated models include XLM-ROBERTa, MultiResCNN, EffectiveCAN, EffectiveCAN (3500), EECAN (Multi-layer attention), and EECAN (Sum-pooling attention). For each model, the chart displays Precision (blue bars), Recall (red bars), and F1 score (green bars). The models show varying performance levels across these metrics, with some being higher in precision while others excel in recall or F1 score. EffectiveCAN and its variations, particularly the 3500 version, demonstrate consistent performance across all three metrics, while EECAN with Multi-layer attention achieves the highest recall. The chart compares each model's strengths and weaknesses in terms of accuracy, memory, and F1 score.

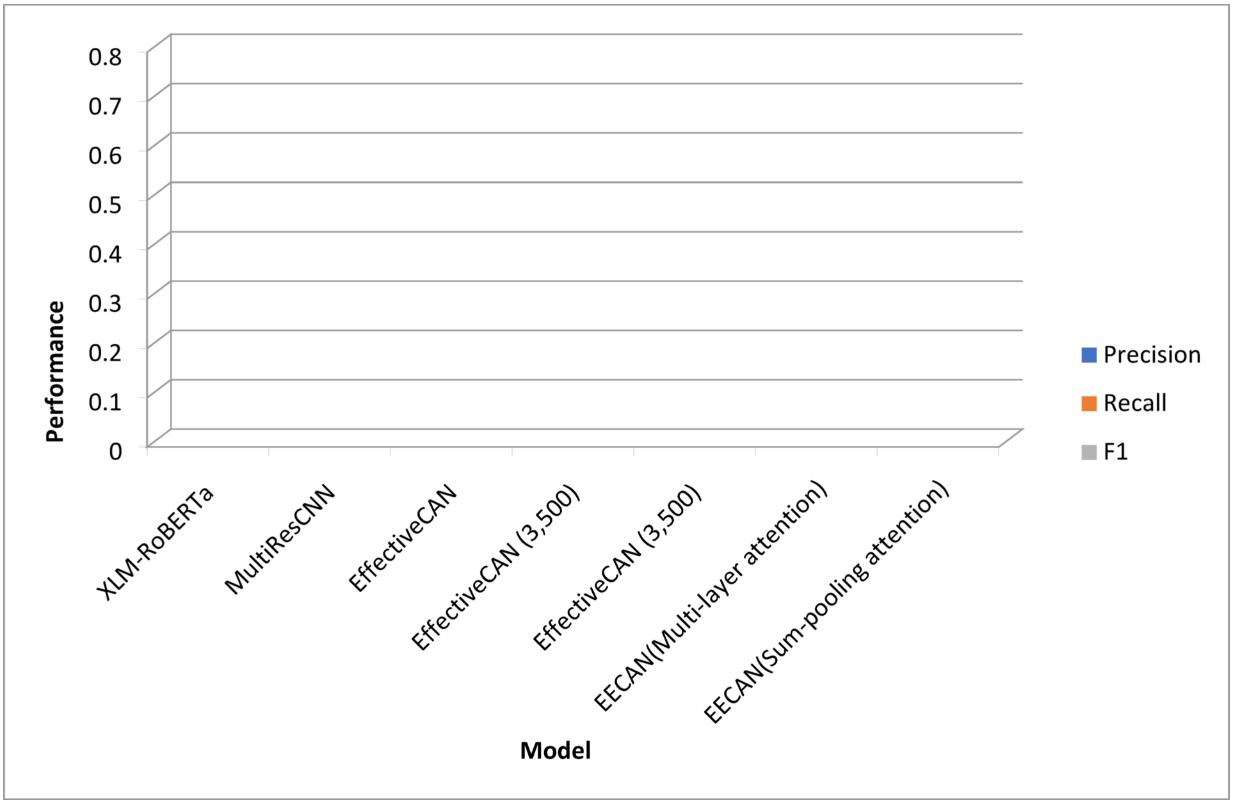


Fig. 8. Results of clinical document classification exhibited by various models using the Dutch dataset.

Configuration	SE-I Module	Attention Mechanism	Focal Loss	AUC	F1-Score	Remarks
Full EECAN (Proposed)	✓	Multi-layer	✓	**99.80%**	**0.725**	Best performance, leveraging SE-I, multi-layer attention, and focal loss
With Sum-Pooling Attention Only	✓	Sum-Pooling	✓	99.70%	0.715	It has a slightly lower performance but is still highly effective for classification
Without SE-I Module	✗	Multi-layer	✓	99.20%	0.685	Reduced feature representation capability, impacting performance
Without Attention Mechanism	✓	✗ (No Attention)	✓	98.70%	0.672	Limited ability to capture label-specific features
Without Focal Loss	✓	Multi-layer	✗	98.90%	0.675	Lower precision and recall for minority classes due to imbalance issues
Without SE-I and Attention Mechanisms	✗	✗ (No Attention)	✓	98.20%	0.651	Significant drop in performance; lacks feature recalibration and focus
Without SE-I, Attention and Focal Loss	✗	✗ (No Attention)	✗	97.80%	0.638	Baseline configuration, struggling to handle label complexity effectively

Table 7. Ablation study of EECAN components on benchmark datasets. Significant values are in bold.

To overcome the challenges of the long-tail distribution of labels, which is frequently observed in clinical document classification, an appropriate model is proposed that utilizes focal loss. It achieves this by adapting the model's focus to the more difficult-to-classify samples by dynamically down-weighting the easier-to-classify examples. Labels are checked in batches, allowing the model to learn more about those samples, especially in datasets with extremely imbalanced classes, where a standard label can overshadow a small one. To assess the model performance on various imbalance levels, we performed experiments where we artificially changed from the original label distributions to create two subsets containing moderate and high imbalance ratios. The results showed that the EECAN model with a focal loss achieved a good classification performance even under heavily imbalanced conditions. Precision, recall, and F1-score for minority classes dropped only a little, indicating the robustness of the model for capturing meaningful features from imbalanced classes. As an illustrative example, Note that an F1-score across all minority classes under the original level of imbalance in the dataset is 0.69; at moderate and high levels of imbalance, this value drops to 0.66 and 0.63, respectively. And while there are some variations from fold to fold, the overall performance of the model—as measured by metrics such as AUROC—was consistently high. Furthermore, by performing these experiments against alternatives such as weighted cross-entropy and class-balanced loss, we assure a conclusive advantage of focal loss in the presence of a class imbalance. This fights the overfitting of majority classes and improves the generalization ability of various label distributions. These observations emphasize the method's adaptability for working with clinically relevant, irregular, and imbalanced label sets.

The ablation studies of the EECAN model, shown in Table 7, emphasize the impact of its modules in their contributions and discuss the efficacy of the SE-I module, attention techniques, and focal loss on benchmark datasets. Overall, the EECAN model with SE-I module, multi-layer attention mechanism, and focal loss outperformed with an AUC of 99.80% (F1-score of 0.725). In conclusion, this configuration effectively allowed the model to capture label-specific features and manage class imbalance issues, among other challenges. Yet its AUC dropped to 99.70% when it used sum-pooling attention instead of a multi-layer attention mechanism, indicating that the latter is still a good alternative. Still, it can capture the complex relationships among the labels less. Leaving out the SE-I module caused a decrease in AUC to 99.20% and a reduction in F1-score, highlighting that the SE-I module is crucial as it improved the features by adjusting channel-wise dependencies. However, all this was with the added attention mechanism entirely removed, resulting in a performance drop to an AUC level of 98.70% and a much poorer ability to learn label-specific features. Likewise, not using focal loss made it difficult to handle imbalanced data distributions (which resulted in lower AUC and minority class precision). The configuration with all components, the SE-I module, attention mechanism, and focal loss outperforms the baseline by a considerable margin, indicating that combining components leads to state-of-the-art performance for multi-label clinical document classification. This demonstrates the optimized classification accuracy and robustness of the SE-I module, attention mechanisms, and focal loss as a synergistic combination in diverse clinical datasets.

In contrast to recent models that have been proposed, EECAN introduces the SE-I module and multi-layer attention mechanisms (including MCB-PS, MCB-ES, and MCB-PS) that work together to update the importance of channel-wise features and to refine label-specific feature extraction. In contrast to global self-attention-based models like BERT or Roberta, EECAN introduces localized feature learning with an effective attention mechanism that leverages computational efficiency and performance gains on clinical datasets. EECAN's SE-I module refines feature representation over hybrid CNN-attention models, yielding higher AUC and F1 scores. Combining these characteristics allows handling long documents with rich and complex label dependencies to perform better than existing models.

We used attention heatmaps to improve the interpretability of our results for clinical professionals and see which parts of the clinical highlights the classifier was paying attention to when making a prediction. Transparency also plays a key role in supporting clinician trust in automated decisions regarding patient data. These visualizations provide insight into the relevance of individual labels to the transcription as a whole. This will be extended in future work with saliency maps and attention rollouts. MIMIC-III-full has a very high label imbalance on the dataset composition level, i.e., the distribution of ICD codes is long-tailed with over 8000 ICD codes. De-identified discharge summaries comprise most datasets, with few other document types included. We use micro and macro-averaged metrics that are resistant to such variation to evaluate since our data has documents that exceed 79 labels. EECAN with multi-layer attention generates the best recall at the cost of moderate precision loss due to more aggressive prediction in the prediction stage, demonstrating the classic precision-recall trade-off. Future studies will also perform statistical significance testing to confirm observed gains in performance. Additionally, EECAN shows competitive accuracy and F1-score compared to existing transformer models but with a substantial improvement in memory efficiency, thus allowing easier deployment into real-world circumstances constrained by hardware and latency.

Statistical significance testing, such as paired *t*-tests or McNemar's test, was not used in this study to compare the performances of models. Nonetheless, the improvements observed over baseline models were consistent across multiple folds and datasets, which provide evidence for the model's effectiveness. Statistical testing will be included in future work to improve the credibility and robustness of the comparative assessment outcomes.

Discussion

The application of artificial intelligence (AI) in the healthcare sector has significantly transformed how clinical data is managed, interpreted, and utilized. AI tools—intense learning models incorporating computer vision, natural language processing, and machine learning—have substantially improved diagnostic accuracy, treatment planning, medical imaging analysis, and patient outcome prediction. In this context, Clinical Document Classification (CDC) is vital in structuring unstructured clinical narratives such as discharge summaries, radiology reports, and laboratory notes. These documents are often lengthy, noisy, and labeled with multiple overlapping conditions, making CDC a complex task that demands models capable of high-level contextual understanding and multi-label learning.

This study introduced EECAN, an Enhanced, Effective Convolutional Attention Network designed for multi-label CDC. EECAN integrates a Squeeze-and-Excitation Inception (SE-I) module for dynamic feature recalibration and employs a dual-attention mechanism—sum-pooling and multi-layer attention—within the EAB-CDC strategy to improve feature focus. The model demonstrated superior performance on benchmark datasets, including MIMIC-III and MIMIC-III-50, achieving AUC scores of 99.70% and 99.80%, respectively. These results outperform several deep learning models for clinical text classification, particularly in handling long documents and diverse labels.

Detailed error analysis revealed that most misclassifications occurred within minority classes or among overlapping ICD codes, reflecting challenges inherent in long-tailed label distributions. In clinical datasets, a small subset of codes appears far more frequently than others, resulting in data imbalance that biases model learning toward standard labels. Incorporating focal loss in EECAN helps mitigate this issue by down-weighting easy examples and emphasizing harder, underrepresented samples, improving recall for minority classes. However, classification performance declined slightly for scarce labels, indicating room for further improvement using class rebalancing techniques or synthetic data augmentation.

Scalability is another crucial factor for deployment in real-world settings. EECAN's modular architecture allows for efficient computation on large clinical datasets, but as dataset size and label space grow, inference latency and memory usage become concerns. Future implementation in production environments—such as EHR systems—must consider hardware constraints, inference speed, and integration with existing data pipelines. Also, handling document diversity across institutions requires robust domain adaptation strategies to maintain model generalization.

Although transformer-based models like BERT, ClinicalBERT, and BioBERT have achieved remarkable results in clinical NLP tasks, they were not included in this study. This choice was due to a focus on developing a lightweight, convolutional, attention-based architecture optimized for efficiency and scalability. Nonetheless, the absence of transformer baselines is a limitation, and future work will incorporate these models for comparison. Evaluating EECAN against state-of-the-art transformer architectures would help clarify the trade-offs between computational cost, interpretability, and performance in classifying long clinical documents. Another critical aspect of real-world adoption is model interpretability. In clinical practice, clinicians must understand why a model makes specific predictions. EECAN's attention mechanisms provide a partial explanation by highlighting informative text regions. Still, future work will explore the integration of Explainable AI (XAI) methods to enhance transparency and trust in model outputs.

Regarding generalizability, EECAN performed well across benchmark datasets; however, variability in language, document formatting, and medical terminologies across clinical institutions may present deployment challenges. Techniques such as transfer learning, domain adaptation, and weak supervision can be explored to address this. Furthermore, testing EECAN on multilingual and cross-institutional datasets (e.g., MIMIC-IV or datasets from non-Western healthcare systems) will provide deeper insights into its robustness and adaptability. Overall, EECAN presents a promising direction for clinical document classification with strong empirical results and a flexible architecture. Future extensions may explore hybrid architectures combining transformers with convolutional modules to leverage the global context modeling of self-attention and the efficiency of convolutional encoders. Real-time deployment in EHR systems, supported by k-fold cross-validation and learning curve analysis, would validate the model's robustness and support its integration into clinical workflows.

Limitations of the study

While the proposed EECAN model has demonstrated strong performance on benchmark datasets, several limitations must be acknowledged. First, clinical document classification tasks often suffer from significant class imbalance, where specific ICD codes or labels appear far more frequently than others. Although focal loss was used to mitigate this issue, performance on rare labels remains challenging and warrants further investigation through advanced resampling or cost-sensitive learning methods.

Second, although EECAN was designed to enhance convolutional attention networks, this study did not incorporate transformer-based models such as BERT, ClinicalBERT, or BioBERT. These models have shown excellent results in clinical NLP tasks due to their ability to model long-range dependencies and contextual relationships. Their exclusion was primarily due to the focus on designing an efficient, lightweight, and scalable architecture. However, the absence of such comparisons limits the generalizability of findings and should be addressed in future work. Third, while EECAN is scalable to large datasets, the model's deployment in real-time clinical environments presents challenges, including inference time, integration with EHR systems, and handling noisy or incomplete records. These operational issues will require further optimization and validation in live healthcare settings. Finally, the current model does not utilize generative adversarial networks (GANs), which hold the potential for synthetic data generation and augmentation—especially valuable in rare or underrepresented classes. Future work could explore GAN-based methods to enrich training data and improve generalization.

Conclusion and future scope

In this study, we proposed the Enhanced Effective Convolutional Attention Network (EECAN) for multi-label clinical document classification, integrating a Squeeze-and-Excitation Inception (SE-Inception) module and an Encoder and Attention-Based Clinical Document Classification (EAB-CDC) strategy. The SE-Inception module enhances feature representation by dynamically recalibrating channel-wise responses. At the same time, EAB-CDC employs sum-pooling and multi-layer attention mechanisms to capture both global and label-specific features. Evaluations on benchmark datasets such as MIMIC-III and MIMIC-III-50 demonstrate that EECAN outperforms existing deep learning models, achieving high AUC values of 99.70% and 99.80%. While these results are promising, the model's reliance on labeled data and its performance across highly imbalanced label distributions present certain limitations. Furthermore, clinical documents with noisy or incomplete content remain challenging for precise classification. These aspects highlight opportunities for further improvement. Incorporating graph-based neural networks could help capture semantic relationships between clinical entities across documents, offering structured reasoning beyond sequence-based models. Similarly, generative adversarial networks (GANs) hold promise in augmenting limited training samples, especially for rare conditions, which can enhance classification robustness and generalizability. To ensure robust performance, applying k-fold cross-validation will allow comprehensive evaluation across multiple data splits, reducing overfitting and improving confidence in the model's reliability. For real-world healthcare deployment, model interpretability and explainability are essential. Future work will explore integrating explainable AI (XAI) methods to provide transparency in model decisions, ensuring that healthcare professionals can trust and act upon the outputs generated by EECAN. Furthermore, while our proposed EECAN model achieves promising results in clinical document classification, future work aims to implement the system within practical healthcare frameworks. A highly effective use case is to embed the model into Electronic Health Record (EHR) systems to automatically classify and tag clinical documents in real-time. Such integration would enable faster access to patient information, supporting diagnosis, treatment planning, and billing—ultimately streamlining clinical workflows. Deploying EECAN in a hospital information system, with access to accurate or de-identified clinical data, will also enable live assessment, illuminate deployment challenges, and demonstrate its value in operational healthcare environments.

Data availability

Data is available with the corresponding author. It will be given on Request.

Received: 17 August 2024; Accepted: 14 April 2025
Published online: 16 May 2025

References

1. Liu, Y., Cheng, H., Klopfer, R., Gormley, M. R. & Schaaf, T. November. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* 5941–5953 (2021).
2. Wang, Y. et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med. Inform. Decis. Making* **19**(1), 1. https://doi.org/10.1186/s12911-018-0723-6 (2019).
3. Kadhim, A. I. Survey on supervised machine learning techniques for automatic text classification. *Artif. Intell. Rev.* https://doi.org/10.1007/s10462-018-09677-1 (2019).
4. Akib Mohi Ud Din, K., Rabani, S. T., Khan, Q. R., Rouf, N. & Mohi Ud Din, M. Machine learning-based approaches for detecting COVID-19 using clinical text data. *Int. J. Inf. Technol.* https://doi.org/10.1007/s41870-020-00495-9 (2020).
5. Kim, D., Seo, D., Cho, S. & Kang, P. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Inf. Sci.* https://doi.org/10.1016/j.ins.2018.10.006 (2018).
6. Raj, J. S. et al. Optimal feature selection based medical image classification using deep learning model in internet of medical things. *IEEE Access* https://doi.org/10.1109/ACCESS.2020.2981337 (2020).
7. Chan, S. et al. Machine learning in dermatology: Current applications, opportunities, and limitations. *Dermatol. Ther.* https://doi.org/10.1007/s13555-020-00372-0 (2020).
8. Huang, L. et al. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J. Biomed. Inform.* **99**, 103291. https://doi.org/10.1016/j.jbi.2019.103291 (2019).
9. Wu, C.-C. et al. Prediction of fatty liver disease using machine learning algorithms. *Comput. Methods Programs Biomed.* **170**, 23–29. https://doi.org/10.1016/j.cmpb.2018.12.032 (2019).
10. Daghistani, T. A. et al. Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *Int. J. Cardiol.* https://doi.org/10.1016/j.ijcard.2019.01.046 (2019).
11. Reddy, G. T., Bhattacharya, S., Siva Ramakrishnan, S., Chowdhary, C. L., Hakak, S., Kaluri, R. & Praveen Kumar Reddy, M. *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)—An Ensemble based Machine Learning model for Diabetic Retinopathy Classification* 1–6. https://doi.org/10.1109/ic-ETITE47903.2020.235 (2020).
12. Ganesan, M. & Sivakumar, N. *IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)—IoT Based Heart Disease Prediction and Diagnosis Model for Healthcare Using Machine Learning Models* 1–5. https://doi.org/10.1109/ICSCAN.2019.8878850 (2019).
13. Piccialli, F., Somma, V. D., Giampaolo, F., Cuomo, S. & Fortino, G. A survey on deep learning in medicine: Why, how and when?. *Inf. Fusion* https://doi.org/10.1016/j.inffus.2020.09.006 (2020).
14. Diwakar, M. et al. Latest trends on heart disease prediction using machine learning and image fusion. *Mater. Today Proc.* https://doi.org/10.1016/j.matpr.2020.09.078 (2020).
15. Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* https://doi.org/10.1007/s42979-021-00592-x (2021).
16. Souri, A. et al. A new machine learning-based healthcare monitoring model for student's condition diagnosis in Internet of Things environment. *Soft Comput.* https://doi.org/10.1007/s00500-020-05003-6 (2020).
17. Ebrahimi, Z., Loni, M., Daneshtalab, M. & Gharehbaghi, A. A review on deep learning methods for ECG arrhythmia classification. *Expert Syst. Appl.* **X**, 100033. https://doi.org/10.1016/j.eswax.2020.100033 (2020).
18. Connor, S., Khoshgoftaar, T. M. & Furht, B. Deep learning applications for COVID-19. *J. Big Data* https://doi.org/10.1186/s40537-020-00392-9 (2021).
19. Sorin, V., Barash, Y., Konen, E. & Klang, E. Deep learning for natural language processing in radiology—Fundamentals and a systematic review. *J. Am. Coll. Radiol.* https://doi.org/10.1016/j.jacr.2019.12.026 (2020).
20. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* https://doi.org/10.1007/s10618-019-00619-1 (2019).
21. Battineni, G., Chintalapudi, N. & Amenta, F. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Inform. Med.* https://doi.org/10.1016/j.imu.2019.100200 (2019).
22. Zhang, J., Xie, Y., Wu, Q. & Xia, Y. Medical image classification using synergic deep learning. *Med. Image Anal.* https://doi.org/10.1016/j.media.2019.02.010 (2019).
23. Wang, X. et al. Inconsistent performance of deep learning models on mammogram classification. *J. Am. Coll. Radiol.* https://doi.org/10.1016/j.jacr.2020.01.006 (2020).
24. Navamani, T. M. Deep learning and parallel computing environment for bioengineering systems. In *Efficient Deep Learning Approaches for Health Informatics* 123–137. https://doi.org/10.1016/B978-0-12-816718-2.00014-2 (2019).
25. Amyar, A., Modzelewski, R., Li, H. & Ruan, S. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. *Comput. Biol. Med.* **126**, 104037. https://doi.org/10.1016/j.compbiomed.2020.104037 (2020).
26. Bohr, A. Artificial intelligence in healthcare. The rise of artificial intelligence in healthcare applications 25–60. https://doi.org/10.1016/B978-0-12-818438-7.00002-2 (2020).
27. Ali, F. et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inf. Fusion* https://doi.org/10.1016/j.inffus.2020.06.008 (2020).
28. Ben-Israel, D. et al. The impact of machine learning on patient care: A systematic review. *Artif. Intell. Med.* https://doi.org/10.1016/j.artmed.2019.101785 (2019).
29. Marshall, I. J. & Wallace, B. C. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst. Rev.* https://doi.org/10.1186/s13643-019-1074-9 (2019).
30. Zhou, T., Ruan, S. & Canu, S. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* https://doi.org/10.1016/j.array.2019.100004 (2019).
31. Solares, A. et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inform.* **101**, 103337. https://doi.org/10.1016/j.jbi.2019.103337 (2020).
32. Sarker, I. H., Kayes, A. S. M. & Watters, P. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *J. Big Data* **6**(1), 57. https://doi.org/10.1186/s40537-019-0219-y (2019).
33. Houssein, E. H., Emam, M. M., Ali, A. A. & Suganthan, P. N. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Syst. Appl.* https://doi.org/10.1016/j.eswa.2020.114161 (2020).

34. Shamshirband, S., Fathi, M., Dehzangi, A., Theodore Chronopoulos, A. & Alinejad-Rokny, H. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *J. Biomed. Inform.* https://doi.org/10.1016/j.jbi.2020.103627 (2020).
35. Zhao, L., Wang, Q., Zou, Q., Zhang, Y. & Chen, Y. Privacy-preserving collaborative deep learning with unreliable participants. *IEEE Trans. Inf. Forensics Secur.* https://doi.org/10.1109/TIFS.2019.2939713 (2019).
36. Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P. & Dhillon, S. K. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med. Inform. Decis. Making* **19**(1), 48. https://doi.org/10.1186/s12911-019-0801-4 (2019).
37. Ashfaq, A., Sant'Anna, A., Lingman, M. & Nowaczyk, S. Readmission prediction using deep learning on electronic health records. *J. Biomed. Inform.* **97**, 103256. https://doi.org/10.1016/j.jbi.2019.103256 (2019).
38. Abdelaziz Ismael, S. A., Mohammed, A. & Hefny, H. An enhanced deep learning approach for brain cancer MRI images classification using residual networks. *Artif. Intell. Med.* https://doi.org/10.1016/j.artmed.2019.101779 (2019).
39. Khan, S. U., Islam, N., Jan, Z., Ud Din, I. & Rodrigues, J. J. P. C. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognit. Lett.* https://doi.org/10.1016/j.patrec.2019.03.022 (2019).
40. Waring, J., Lindvall, C. & Umeton, R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Med.* https://doi.org/10.1016/j.artmed.2020.101822 (2020).
41. Mujtaba, G. et al. Clinical text classification research trends: Systematic literature review and open issues. *Expert Syst. Appl.* **116**, 494–520. https://doi.org/10.1016/j.eswa.2018.09.034 (2019).
42. Suárez-Paniagua, V., Rivera Zavala, R. M., Segura-Bedmar, I. & Martínez, P. A two-stage deep learning approach for extracting entities and relationships from medical texts. *J. Biomed. Inform.* **99**, 103285. https://doi.org/10.1016/j.jbi.2019.103285 (2019).
43. Gargiulo, F., Silvestri, S., Ciampi, M. & De Pietro, G. Deep neural network for hierarchical extreme multi-label text classification. *Appl. Soft Comput.* https://doi.org/10.1016/j.asoc.2019.03.041 (2019).
44. Topaz, M. et al. sMining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *J. Biomed. Inform.* https://doi.org/10.1016/j.jbi.2019.103103 (2019).
45. Yadav, S. S. & Jadhav, S. M. Deep convolutional neural network based medical image classification for disease diagnosis. *J. Big Data* **6**(1), 113. https://doi.org/10.1186/s40537-019-0276-2 (2019).
46. Garg, R., Oh, E., Naidech, A., Kording, K. & Prabhakaran, S. Automating ischemic stroke subtype classification using machine learning and natural language processing. *J. Stroke Cerebrovasc. Dis.* https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.02.004 (2019).
47. Juhn, Y. & Liu, H. Natural language processing to advance EHR-based clinical research in allergy, asthma, and immunology. *J. Allergy Clin. Immunol.* https://doi.org/10.1016/j.jaci.2019.12.897 (2019).
48. Dreisbach, C., Koleck, T. A., Bourne, P. E. & Bakken, S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int. J. Med. Inform.* https://doi.org/10.1016/j.ijmedinf.2019.02.008 (2019).
49. Zhang, F., Fleyeh, H., Wang, X. & Lu, M. Construction site accident analysis using text mining and natural language processing techniques. *Autom. Constr.* **99**, 238–248. https://doi.org/10.1016/j.autcon.2018.12.016 (2019).
50. Kraljevic, Z. et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artif. Intell. Med.* https://doi.org/10.1016/j.artmed.2021.102083 (2021).
51. Kormilitzin, A., Vaci, N., Liu, Q. & Nevado-Holgado, A. Med7: A transferable clinical natural language processing model for electronic health records. *Artif. Intell. Med.* https://doi.org/10.1016/j.artmed.2021.102086 (2021).
52. Locke, S. et al. Natural language processing in medicine: A review. *Trends Anaesth. Crit. Care.* https://doi.org/10.1016/j.tacc.2021.02.007 (2021).
53. Pandey, B., Pandey, D. K., Mishra, B. P. & Rhmann, W. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challen. *Elsevier* **34**(8), 5083–5099. https://doi.org/10.1016/j.jksuci.2021.01.007 (2022).
54. Casey, A. et al. A systematic review of natural language processing applied to radiology reports. *BMC Med. Inform. Decis. Making.* https://doi.org/10.1186/s12911-021-01533-7 (2021).
55. Gao, S. et al. Limitations of transformers on clinical text classification. *IEEE J. Biomed. Health Inform.* https://doi.org/10.1109/jbhi.2021.3062322 (2021).
56. Li, X. et al. A hybrid medical text classification framework: Integrating attentive rule construction and neural network. *Neurocomputing* https://doi.org/10.1016/j.neucom.2021.02.069 (2021).
57. Alhogail, A. & Alsabih, A. Applying machine learning and natural language processing to detect phishing email. *Comput. Secur.* https://doi.org/10.1016/j.cose.2021.102414 (2021).
58. Barber, E. L., Garg, R., Persenaire, C. & Simon, M. Natural language processing with machine learning to predict outcomes after ovarian cancer surgery. *Gynecol. Oncol.* https://doi.org/10.1016/j.ygyno.2020.10.004 (2020).
59. Alexander, T., Masharsky, S. & Zitnik, M. Comparison of BERT implementations for natural language processing of narrative medical documents. *Elsevier* **36**, 1–7. https://doi.org/10.1016/j.imu.2022.101139 (2023).
60. Fernandes, M. B. et al. Classification of neurologic outcomes from medical notes using natural language processing. *Elsevier* **214**, 1–10. https://doi.org/10.1016/j.eswa.2022.119171 (2023).
61. Han, S. et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural. *Elsevier* **127**, 1–11. https://doi.org/10.1016/j.jbi.2021.103984 (2022).
62. Zhou, B., Yang, G., Shi, Z. & Shaodan, M. Natural language processing for smart healthcare. *IEEE* https://doi.org/10.1109/RBME.2022.3210270 (2022).
63. Khanbhai, M. et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. *Elsevier* **157**, 1–7. https://doi.org/10.1016/j.ijmedinf.2021.104642 (2022).
64. Lavanya, P. M. & 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)*. https://doi.org/10.1109/ICSPC51351.2021.9451752 (2021).
65. Richard, E. & Reddy, B. Text classification for clinical trial operations: Evaluation and comparison of natural language processing techniques. *Ther. Innov. Regul. Sci.* https://doi.org/10.1007/s43441-020-00236-x (2020).
66. López-Úbeda, P. et al. Automatic medical protocol classification using machine learning approaches. *Comput. Methods Programs Biomed.* https://doi.org/10.1016/j.cmpb.2021.105939 (2021).
67. Hammami, L. et al. Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing techniques: A rule-based approach. *J. Biomed. Inform.* https://doi.org/10.1016/j.jbi.2021.103712 (2021).
68. Kulshrestha, S. et al. Prediction of severe chest injury using natural language processing from the electronic health record. *Injury* https://doi.org/10.1016/j.injury.2020.10.094 (2020).
69. Borjali, A. et al. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation. *Comput. Biol. Med.* **129**, 104140. https://doi.org/10.1016/j.compbiomed.2020.104140 (2021).
70. Johnson, S. A. et al. A comparison of natural language processing to ICD-10 codes for identification and characterization of pulmonary embolism. *Thromb. Res.* **203**, 190–195. https://doi.org/10.1016/j.thromre (2021).
71. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7132–7141 (2019).

72. Esraa Hassan, A. S. & Elbedwehy, S. Knowledge distillation model for acute lymphoblastic leukemia detection: Exploring the impact of nesterov-accelerated adaptive moment estimation optimizer. *Biomed. Signal Process. Control* **94**, 106246. https://doi.org/10.1016/j.bspc.2024.106246 (2024).
73. Saber, A. et al. An optimized ensemble model based on meta-heuristic algorithms for effective detection and classification of breast tumors. *Neural Comput. Appl.* **37**, 4881–4894. https://doi.org/10.1007/s00521-024-10719-9 (2025).
74. Elbedwehy, S. et al. Integrating neural networks with advanced optimization techniques for accurate kidney disease diagnosis. *Sci. Rep.* **14**, 21740. https://doi.org/10.1038/s41598-024-71410-6 (2024).
75. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
76. Hou, L. et al. Normalization helps training of quantized lstm. *Adv. Neural Inf. Process. Syst.* **32**, 1–11 (2019).
77. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1–9. https://doi.org/10.1109/cvpr.2016.90 (2016).
78. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J. & Eisenstein, J. Explainable prediction of medical codes from clinical text. arXiv preprint arXiv:1802.05695, 1–11 (2018).
79. Sánchez, J., Perronnin, F., Mensink, T. & Verbeek, J. Image classification with the fisher vector: Theory and practice. *Int. J. Comput. Vis.* **105**, 222–245 (2013).
80. Shen, L. et al. Multi-level discriminative dictionary learning with application to large scale image classification. *IEEE Trans. Image Process.* **24**(10), 3109–3123. https://doi.org/10.1109/tip.2015.2438548 (2015).
81. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning* (ICML-10) 807–814 (2010).
82. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J. & Eisenstein, J. Explainable prediction of medical codes from clinical text 1–11. arXiv preprint arXiv:1802.05695 (2018).
83. Chalkidis, I., Fergadiotis, M., Kotitsas, S., Malakasiotis, P., Aletras, N. & Androutsopoulos, I. An empirical study on large-scale multi-label text classification including few and zero-shot labels 1–13. arXiv preprint arXiv:2010.01653 (2020).
84. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. & Androutsopoulos, I. Extreme multi-label legal text classification: A case study in EU legislation 1–10. arXiv preprint arXiv:1905.10892 (2019).
85. Shi, H., Xie, P., Hu, Z., Zhang, M. & Xing, E. P. Towards automated ICD coding using deep learning 1–11. arXiv preprint arXiv:1711.04075 (2017).
86. Liaw, A. & Wiener, M. Classification and Regression by RandomForest. *R News* **2**(3), 18–22 (2002).
87. Vu, X. V., Nguyen, H. T., Tran, T. & Phung, D. Label Attention Model for ICD Coding. In *Proceedings of the 19th International Conference on Artificial Intelligence in Medicine (AIME 2020). Springer, Lecture Notes in Computer Science*, vol 12299, p 290–300 (2020).

Author contributions

All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by M. Venkata Krishna Reddy, Dr. L. Raghavendar Raju, Dr. Kashi Sai Prasad, Dr. D. Anitha Kumari, Vadlamani Veerabhadram, and Nagendar Yamsani. M. Venkata Krishna Reddy wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.V.K.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.