Journal of Electronics, Electromedical Engineering, and Medical Informatics
Vol. 7, No. 3, July 2025, pp: 800-816 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Automated ICD Medical Code Generation for Radiology Reports using BioClinicalBERT with Multi-Head Attention Network

Sasikala D[1], Sarrvesh N[1], Sabarinath J[1], Theetchenya S[2], and Kalavathi S[3]
[1] Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India
[2] Department of Computer Science and Engineering, Sona College of Technology, Salem, India
[3] Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Chennai, India

Corresponding author: Sasikala D. (e-mail: d_sasikala@ch.amrita.edu), **Email Author(s)**: Sarrvesh N (e-mail:nsarrvesh6710@gmail.com), Sabarinath J (e-mail:sabarinath.jsn@gmail.com), Theetchenya S (e-mail:theetchenya@gmail.com), Kalavathi S(e-mail:kalavathi@svce.ac.in)

Abstract **International Classification of Diseases (ICD) coding plays a pivotal role in healthcare systems with its provision of a standard method for classifying medical diagnoses, treatments, and procedures. However, the process of manually applying ICD codes to clinical records is both time-consuming and error-prone, particularly considering the large magnitude of medical terminologies and the periodic changes to the coding system. This work introduces a Hierarchical Multi-Head Attention Network (HMHAN) that aims to automate ICD coding using domain-related embeddings with an attention mechanism. The proposed method uses BioClinicalBERT for feature extraction from clinical text and then a two-level attention mechanism to learn hierarchical dependencies between labels. BioClinicalBERT is pre-trained on large biomedical and clinical corpora that enable it to capture complex contextual relationships specific to medical language more effectively. The multi-head attention mechanism enables the model to focus on different parts of the input text simultaneously, learning intricate associations between medical terms and corresponding ICD codes at various levels. This method uses SMOTE (Synthetic Minority Oversampling Technique) based multi-label resampling to solve class imbalance. SMOTE generates synthetic examples for underrepresented classes, allowing the model to learn better from imbalanced data without overfitting. For this work, MIMIC-IV dataset of de-identified radiology reports and corresponding ICD codes are used. The performance of the model is assessed with F1 score, Hamming loss, and ROC-AUC metrics. Results obtained from the model with an F1 score of 0.91, Hamming loss of 0.07, and ROC-AUC of 0.92 show promising research directions to automate the ICD coding process. This system will improve the effectiveness of healthcare workflows by automating ICD code generation for advanced clinical care.**

Keywords **Automated ICD coding; Radiology reports; MIMIC-IV; Hierarchical Multi-Head Attention Network; BioClinicalBERT; Health Informatics.**

I. Introduction

The increasing volume of electronic health records (EHRs) has necessitated the development of automated medical coding systems to streamline clinical documentation and billing processes. Patient diagnoses and procedures get classified through the International Classification of Diseases (ICD) coding system to support healthcare analytics research and reimbursement as well as epidemiological studies [1]. Healthcare professionals spend considerable time on manual ICD coding, and sometimes it leads to errors because it needs specialized knowledge on medical field. Machine Learning (ML) and Deep Learning (DL) techniques have been adopted to automate ICD coding procedures according to recent literature [2]. Natural language processing (NLP) and deep neural networks (DNNs) recently improved automation of ICD coding system. The research demonstrates that deep learning models including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) achieve better results in extraction of ICD code from clinical text documents [3]. Deep learning models have obstacles such as extended text relationships along with unequal distribution of labels and difficulties, with interpretation [4]. The proposed hierarchical multi-head attention networks (HMHAN) represent an effective solution to address these problems by exploiting multiple attention layers which enhance contextual understanding.

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

Several studies have explored deep learning based ICD coding systems. Masud et al. [5] developed a CNN based model that achieved notable precision and recall in ICD code prediction. Oberste et al. [6] proposed an NLP driven ML model for outpatient billing, improving reimbursement accuracy. Teng et al. [7] provided a comprehensive review of deep learning applications in ICD coding, emphasizing the importance of multi label classification. Vu et al. [8] introduced a label attention model to handle imbalanced data distributions and enhance ICD coding accuracy. Similarly, Kim et al. [9] utilized partition based label attention to improve token level representation learning. Transformer based models have also been explored for extreme multi label classification tasks, achieving state-of-the-art performance in automated ICD coding [10].

Despite these advancements, challenges remain in achieving high accuracy, scalability, and interpretability in automated ICD coding systems. This work contributes to the existing knowledge by introducing a novel automated ICD coding framework based on the Hierarchical Multi-Head Attention Network (HMHAN). The proposed approach uniquely integrates hierarchical label-wise attention mechanisms, not only to improve the accuracy of code prediction but also to enhance the transparency of the model's decision making process. By explicitly modeling the hierarchical relationships within the ICD coding system and employing multi-head attention at different levels, this work aims to refine the ICD coding process, reduce errors, improve efficiency in clinical documentation, and offer insights into the model's predictions, thereby addressing a key limitation of current deep learning approaches.

II. Literature Survey

Automated ICD coding has become an essential focus in healthcare, given its potential to reduce manual coding errors, streamline hospital workflows, and improve overall coding accuracy. The complex nature of medical terminologies along with frequent updates in the ICD coding system makes automated solutions particularly valuable. Researchers have turned to address the rare code prediction, and interpretability using machine learning and deep learning methods to address the core challenges of multi label classification.

Yuan et al. [11] systematically reviewed 118 studies, highlighting how AutoML not only reduces the barrier for healthcare professionals with limited ML expertise but also enhances model development efficiency across diverse data modalities, including images, text, and genomic data. However, the inherent black-box nature of many AutoML systems raises concerns regarding their interpretability, which is essential in clinical contexts where model transparency is crucial for trust and regulatory compliance. To address this, the integration of interpretation methods such as feature importance analysis, intrinsically interpretable models, and rule extraction has been proposed to improve model transparency and foster clinical adoption. The review emphasizes the need for future research to further automate interpretability processes and to support multi modal data and foundation models to bridge the gap between technical innovation and real world implementation in healthcare.

Rohil and Magotra [12] conducted an exploratory study on automatic text summarization (ATS) in the biomedical and healthcare domains. Their study analyzed various ATS approaches, including extractive and abstractive methods and evaluated their effectiveness in summarizing clinical records, electronic health records (EHRs) and biomedical literature. By comparing different summarization techniques, they provided insights into the strengths and limitations of ATS in handling complex medical texts. Their research highlighted that extractive methods are effective in retaining key information from documents, while abstractive methods offer more concise and human readable summaries. The findings are relevant to ICD coding, as text summarization techniques can be adapted to process clinical notes efficiently and generate meaningful summaries that assist in medical code assignment.

Jayanth et al. [13] investigated the use of XLM-RoBERTa for intent recognition in natural language understanding (NLU) applications. By leveraging the multilingual capabilities of XLM-RoBERTa, the study focused on enhancing the accuracy of intent detection across different languages. The research showed that XLM-RoBERTa outperformed traditional models in understanding diverse language structures and handling variations in sentence phrasing. The model's robustness in intent recognition has implications for ICD coding, where understanding the subtle nuances in clinical notes is crucial for accurate code assignment. This study underscores the adaptability of transformer models to complex NLP tasks beyond standard language domains.

Ponthongmak et al. [14] explored automated ICD-10 coding using deep learning techniques applied to discharge summaries. Their research leveraged natural language processing (NLP) and various models, including CNN-PubMedBERT. Results demonstrated that CNN-PubMedBERT outperformed traditional approaches, through PLM-ICD, incorporating label wise attention and RoBERTa-PubMed embeddings exhibited superior performance overall.

Wang et al. [15] introduced ICDXML, enhancing ICD coding through probabilistic label trees and dynamic semantic representations. The approach tackled long

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

tailed ICD codes by leveraging hierarchical structures and integrating semantic representations, achieving notable improvements in precision and recall rates, especially for rare ICD codes. Zhao et al. [16] applied transformer based models with attention mechanisms to identify relevant sections of clinical notes for coronary heart disease diagnosis. These models emphasized important parts of the text, leading to significant gains in recall and F1-scores, particularly effective for multi label classification of complex conditions.

Wu et al. [17] introduced a hyperbolic graph convolutional network combined with ensemble methods to improve ICD code assignment, utilizing contrastive learning to differentiate similar codes. Similarly, Bhutto et al. [18] developed a Lambda Scaled Attention based model integrating CNNs, LSTMs, and attention mechanisms, significantly improving accuracy by focusing on clinically relevant text portions. Chen et al. [19] presented an innovative approach using deep semantic matching based on analogical reasoning. Their model integrated BERT based word representation, BiLSTM context representation, and multi perspective matching layers, achieving state-of-the-art performance with 0.986 accuracy and 0.981 F1-score. The approach compared uncoded diagnoses with previously coded records rather than directly with ICD-10 terminology.

Zhao et al. [20] applied contrastive learning with transformer and CNN architectures, focusing on variations in ICD code representations across different medical records. This improved accuracy, sensitivity, and specificity, especially in cases involving multiple diagnoses. Coutinho and Martins [21] investigated Transformer-based models for ICD-10 coding of Portuguese death certificates. Their BERT-based model leveraged domain-specific pre-training and fine-tuning strategies, outperforming traditional approaches, particularly for short clinical narratives. Chomutare et al. [22] leveraged a Swedish language model (KB-BERT) with fuzzy logic to enhance prediction of rare ICD codes, effectively handling ambiguous terms in clinical notes and improving precision and recall.

Shuai et al. [23] compared feature extraction methods for automated coding, finding that fine-tuned BERT networks performed best for frequent codes, while bag-of-words outperformed deep learning methods when datasets included both frequent and infrequent codes. Bhutto et al. [24] introduced the Deep Recurrent Convolutional Neural Network with Transfer Learning through Pre-trained Embeddings (DRCNNTLe), leveraging pre-trained word embeddings to enhance text representations for liver transplant patients. Chen et al. [25] integrated embeddings from transformers, global vectors, word2vec, and a single head attention recurrent neural network within a GRU framework, achieving F1-scores of 0.715 for ICD-10 Clinical Modification codes. Their web service enhanced coders' accuracy but did not reduce manual coding time.

Diao et al. [26] presented a clinically interpretable model for cardiovascular diseases in China, employing sequential grouping features with Light Gradient Boosting Machine classifiers. The model achieved 95.2% accuracy and 88.3% macro-averaged F1-score, with SHapley Additive exPlanations enhancing interpretability. Makohon and Li [27] addressed challenges of abbreviation normalization and misspelled words, introducing a hierarchical approach converting lower level ICD codes into respective ICD chapters. Their Hierarchical Attention Network with GRU achieved the highest F1-score in cross validation. Luo et al. [28] presented the Fusion model, using attention-based soft-pooling to condense sparse information into meaningful features, outperforming several state-of-the-art models in multiple evaluation metrics.

Chraibi et al. [29] proposed a deep learning framework for French electronic health records, achieving 83% average accuracy across 346 diagnosis codes. Cao et al. [30] introduced Clinical-Coder for Chinese clinical notes, using a Dilated Convolutional Attention Network with N-gram Matching to enhance interpretability. Zhang et al. [31] developed BERT-XML, trained on over 5 million EHR notes with domain-specific vocabulary and extended sequence length, significantly improving prediction accuracy for 2,292 ICD-10 codes. Huang et al. [32] found GRU-based models performed best for ICD-9 code prediction, achieving an F1-score of 0.6957.

Recent studies have employed extractive summarization with PubMedBERT and BioBERT, demonstrating superior performance in medical text analysis [33][34]. Advancements with GPT-3 and PEGASUS have improved named entity recognition for automated coding. Multi language applications show significant improvements in classification accuracy, while fuzzy logic and knowledge based systems enhance interpretability and robustness in clinical settings.

This review highlights the advancements in automated ICD coding using various deep learning and machine learning approaches, emphasizing their role in improving accuracy, handling rare codes, and ensuring interpretability in clinical settings.

III. Materials and Method

A. Dataset

This project works with data from the MIMIC-IV (Medical Information Mart for Intensive Care IV)

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

database which contains de-identified health information about more than 60,000 critical care patients [35]. The research examines radiology notes found in MIMIC-IV since this free text clinical documentation holds essential patient assessment information about diagnoses and health history. Deriving predictions of ICD-9 and ICD-10 codes represents the objective while processing radiology reports together with additional patient data. This research examined the top ten ICD codes Fig. 1 present in a dataset which holds more than 70,000 records. Table 1 represents the dataset characteristics for the TOP10 ICD Code used for this work. To access MIMIC-IV researchers must first finish the Protecting Human Research Participants course from the National Institutes of Health (NIH) and sign the Data Use Agreement (DUA) compliance statement to protect confidential healthcare information.

This analysis examines MIMIC-IV's radiology information by integrating unstructured and structured data to create predictive models that automate ICD coding procedures [36]. The radiology.csv dataset contains free text unstructured Radiology Notes that document medical imaging results including X-rays, CT scans, MRIs and ultrasounds along with radiology interpretations from radiologists in their indications, findings and impressions sections. Radiology Detail (radiology_detail.csv) delivers structured metadata about radiology reports that enhances predictive modelling features by listing imaging modality and study type and body part examination.

Both the ICD Codes datasets (diagnoses_icd.csv and procedures_icd.csv) contain ICD-9 and ICD-10 codes which function as model target labels but require a multi label classification technique because of their hierarchical structure. The patient demographic data in patients.csv allows derivation of calculated features by combining date of birth with date of death to obtain patient age. Multiple structured and unstructured datasets provided by radiology notes, metadata, ICD codes and demographic information create an extensive data set for developing sophisticated machine learning systems [37]. Healthcare narratives contained in unstructured text provide detailed clinical data alignment with structured metadata features to develop complete predictive models for ICD code automation.

Table 1. Dataset Characteristics of Top 10 ICD codes of radiology reports from MIMIC-IV

Description	Value
Number of Samples	73461
Number of Columns	26
Average Age	43
Gender (Male/Female)	42027 / 31434
Number of ICD Codes	10

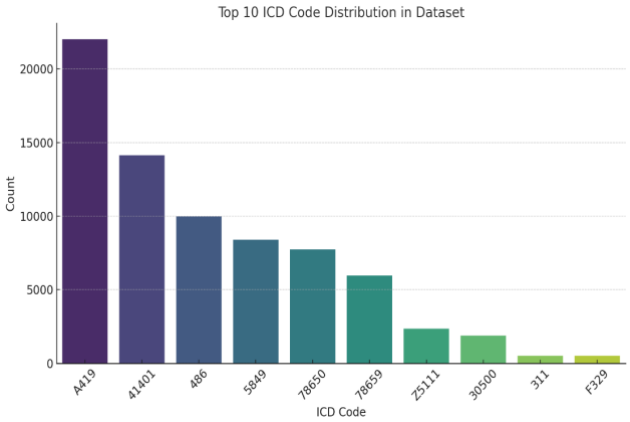


Fig. 1. Distribution of the Top 10 ICD Codes of radiology reports from MIMIC-IV

B. Data Preprocessing

The raw clinical text data undergoes a series of preprocessing steps to ensure high quality input for the machine learning models. These steps were designed to handle the unique challenges of medical text, such as negation, complex terminology, and inconsistent formatting.

Negation plays a critical role in clinical diagnosis. For example, the phrases "pneumonia present" and "no evidence of pneumonia" have opposite meanings but could be treated similarly by naive models. To address this, we employed the NegEx algorithm, which identifies and tags negated terms by appending a "NOT_" prefix (e.g., "NOT_pneumonia"). This ensures that negated conditions are not mistakenly considered as actual diagnoses during feature extraction. By distinguishing affirmed from negated terms, this step improves both precision and recall in ICD code prediction.

We utilized subword based tokenization compatible with BioClinicalBERT. Tokenization splits the text into meaningful units (tokens), such as transforming "lung infection detected" into ["lung", "infection", "detect", "##ed"] [38]. This is particularly important for rare or compound medical terms, which may not appear frequently in the model's vocabulary. Tokenization ensures that even unseen or partially known words are

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

represented effectively, improving the model's ability to capture semantic nuance in medical text.

Clinical notes often contain inconsistent capitalization, misspellings, punctuations, and irrelevant numerical data. Text cleaning includes converting text to lowercase, removing special characters, standardizing abbreviations, and filtering out stopwords. For example, "Patient has a fever of 102°F" is cleaned to "patient has fever". This reduces noise and variability in the input, enhancing the stability and generalizability of the learned embeddings.

We standardized numerical features (e.g., age) using z-score normalization and applied one-hot encoding to categorical features like gender or admission type. This allows the model to treat each feature on a comparable scale, ensuring balanced gradient updates during training. Transformer models like BioClinicalBERT require fixed length input sequences. We padded shorter sequences using the [PAD] token and truncated longer sequences to a maximum of 512 tokens. This uniformity enables efficient batch processing while preserving as much clinical context as possible within the allowable input size.

C. BioClinicalBERT for Embeddings Generation

BioClinicalBERT served as the domain specific BERT variant to extract features from medical content in clinical documents. Unlike general purpose transformer models such as BERT or RoBERTa, BioClinicalBERT is pre-trained specifically on large-scale biomedical corpora including MIMIC-III and clinical notes from electronic health records. This pre-training enables it to better understand domain specific terminologies, contextual cues, and medical jargon that are prevalent in healthcare data. Table 2 specifies BioClinicalBERT parameters. One of the primary reasons for selecting BioClinicalBERT over other models is its proven effectiveness in handling clinical language. For instance, general BERT-based models may struggle with medical terms like "atelectasis", "pneumothorax", or contextual phrases such as "rule out MI" which have specific meanings in clinical settings. BioClinicalBERT, however, can capture these nuances more accurately due to its training on similar language.

In the context of ICD coding, where fine grained distinctions between conditions directly influence code assignments, capturing these subtle contextual meanings is crucial. BioClinicalBERT generates contextual embeddings that reflect both the surrounding text and the semantics of the clinical terms, making it highly suitable for downstream tasks like multi label classification of ICD codes. Moreover, compared to other biomedical models like BioBERT or ClinicalBERT individually, BioClinicalBERT combines the strengths of both biomedical and clinical corpora, offering a broader understanding of both general biomedical literature and patient centric records. Its architecture and vocabulary are fine-tuned for clinical reasoning tasks, which contributes to improved performance in extracting meaningful patterns from unstructured radiology reports. The embeddings generated by BioClinicalBERT serve as input features to the Hierarchical Multi-Head Attention Network (HMHAN). These embeddings capture the rich linguistic and semantic information present in clinical narratives, enabling the HMHAN to leverage deep contextual understanding for accurate multi label classification. By utilizing BioClinicalBERT, the model can effectively process unstructured clinical text, such as radiology reports, and extract meaningful features that align with the hierarchical structure of ICD codes.

D. Handling Class Imbalance

A major challenge exists for automated ICD coding because a small number of ICD codes appears frequently in the data and many ICD codes remain rare. The Synthetic Minority Over Sampling Technique (SMOTE) was modified to work with multi label data to solve this matter [39][40]. SMOTE creates new representative examples through the process of interpolation between current minority class observations thus achieving class balance between common and uncommon ICD codes. By using this technique, the model avoids indiscriminate preference for dominant codes allowing it to detect both infrequent and important medical classification codes better. Table 3 represents the impact of SMOTE in this work. The training process used additional weight calculated losses which included focal loss functioning with binary cross entropy to enhance emphasis on 'hard to find' diagnosis codes. These methods build stronger model capabilities which extend to predict ICD codes across all possible categories.

Table 2. BioClinicalBERT Parameters

Parameter	Value
Model Type	BioClinicalBERT
Number of Layers	12
Hidden Size	768
Number of Attention Heads	12
Max Sequence Length	512 tokens
Total Parameters	~110 million
Activation Function	GELU
Pre-training Dataset	MIMIC-III

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

Table 3. Impact of SMOTE in dataset

Aspect	Without SMOTE	With SMOTE
Class Distribution	Imbalanced (Few samples for rare classes)	Balanced (Synthetic samples for rare classes)
Impact on Model	Biases toward frequent classes	Provides fairer learning across all classes
Rare Class Detection	Poor (Rare cases often misclassified)	Improved (Better representation of rare classes)
Overfitting Risk	Lower (Trains only on real data)	Moderate (Synthetic samples may introduce noise)
Accuracy vs. Recall	Higher accuracy but low recall for minority classes	Balanced accuracy and improved recall

The combination of SMOTE with class weighted loss functions served to remedy the problems from class imbalance in the data. Binary cross-entropy with focal loss function was implemented for training to boost the emphasis on underreported codes. With focal loss a model can regulate the loss value from correctly predicted cases by applying a gamma parameter which lets the model focus on harder minority class samples. When SMOTE combines with focal loss it prevents the model from excessively weighting dominant classes to become more successful at identifying rare and crucial ICD codes. The technique validations were completed through intensive experimental testing. The model delivers impressive advancements of recall and F1-score metrics regarding rare ICD codes by using label balancing techniques together with underrepresented code prioritization. The enhanced predictive model demonstrates better performance in practice settings because it effectively handles rare medical diagnostic codes that represent complex or urgent clinical situations.

E. Comparative Evaluation with Progressive Baselines

To evaluate the performance and significance of the proposed Hierarchical Multi-Head Attention Network (HMHAN), a series of progressively complex baseline models were implemented and benchmarked under identical experimental conditions. These comparative baselines allow for a transparent analysis of the architectural improvements introduced in this study. The three models BioClinicalBERT without attention, BioClinicalBERT [41] with a custom attention mechanism, and HMHAN represent an evolutionary trajectory in model design aimed at improving automated ICD code prediction.

1. BioClinicalBERT without Attention Mechanism

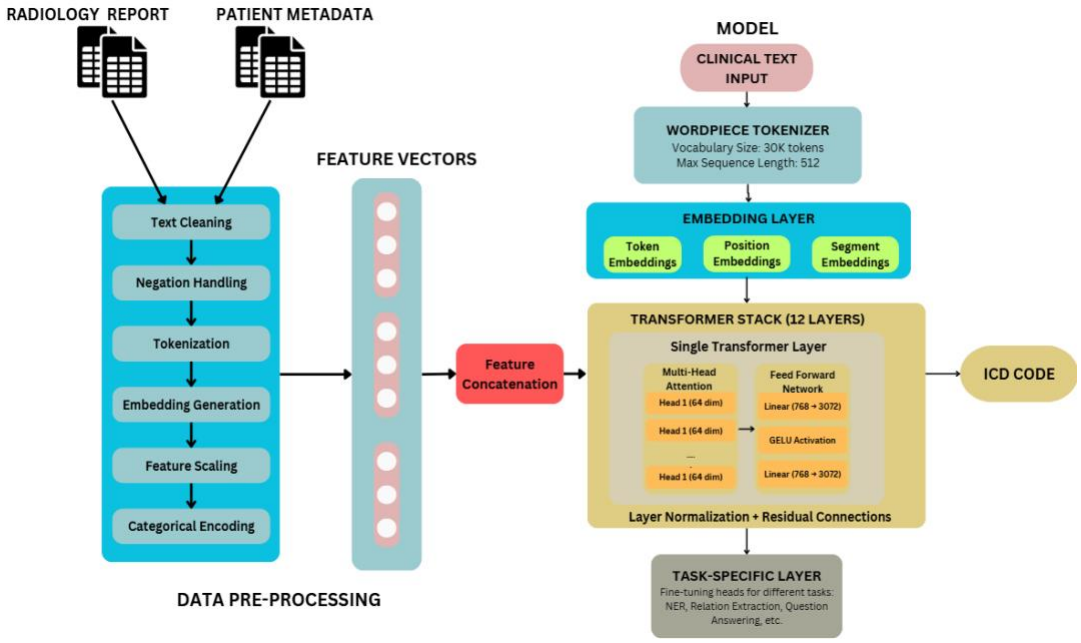


Fig 2. Architecture of BioClinicalBERT without Attention.

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

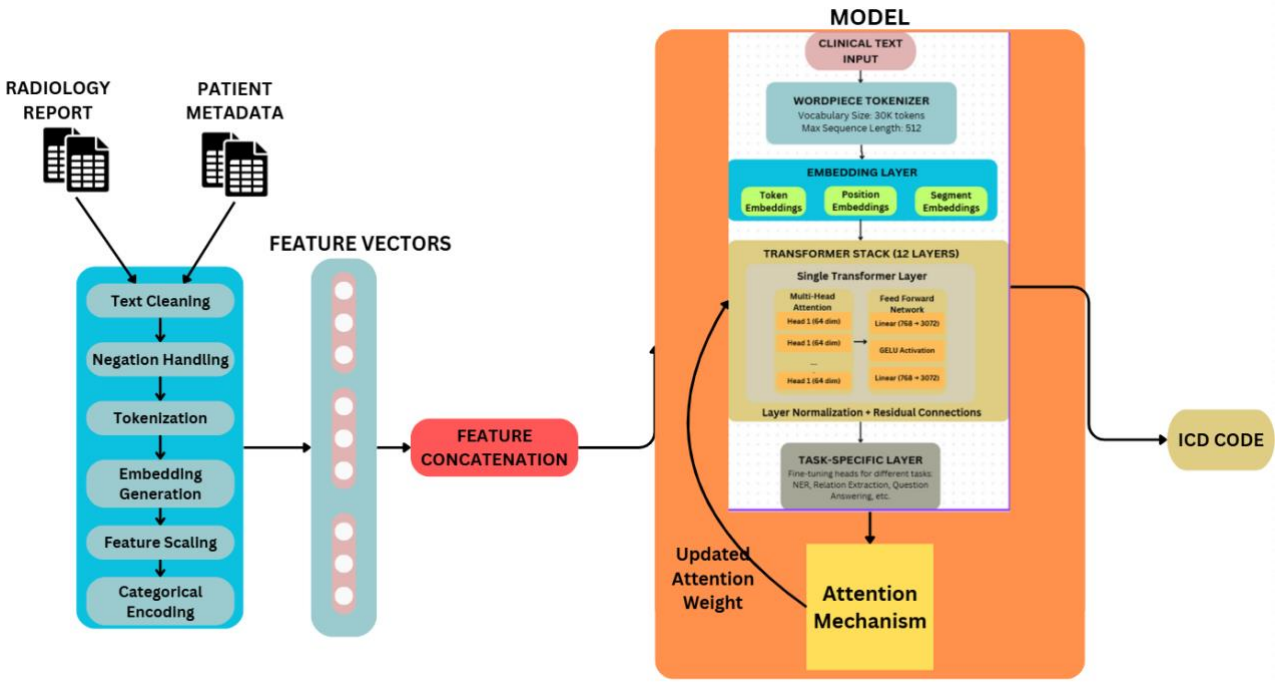


Fig 3. Architecture of BioClinicalBERT with Attention.

As an initial baseline, BioClinicalBERT was utilized purely as a contextual feature extractor. The token level embeddings generated by the model were passed directly through fully connected layers to perform multi label classification without any intermediate attention mechanism. Despite achieving high precision (0.82), this configuration suffered from poor recall (0.28), indicating a tendency to under predict relevant codes. The F1-score of 0.42 reflected the model's imbalance in capturing both true positives and false negatives effectively. The lack of an attention mechanism limited the model's ability to dynamically emphasize important features from the clinical text. The architecture of this baseline is illustrated in Fig. 2,where BioClinicalBERT embeddings flow directly into dense layers, emphasizing its simplicity and lack of contextual focus enhancement.

2. BioClinicalBERT with Custom Attention Mechanism

To improve upon the limitations observed in the previous model, a custom attention layer was added atop the BioClinicalBERT embeddings. This layer was designed to assign greater weight to clinically relevant words and phrases within the radiology reports, thereby enhancing the model's focus on informative regions. The inclusion of attention resulted in an improved recall (0.50), but came with a reduction in precision (0.30) due to increased false positive predictions. The overall F1-score decreased to 0.37, suggesting that while sensitivity improved, specificity was compromised. The architecture is depicted in Fig. 3, showcasing how the attention mechanism selectively amplifies key token embeddings before classification, offering a more dynamic and interpretable learning pipeline.

3. Proposed Model: Hierarchical Multi-Head Attention Network (HMHAN)

To further balance the trade off between precision and recall and address the hierarchical nature of ICD codes, the Hierarchical Multi-Head Attention Network (HMHAN) was introduced. This architecture employs dual level attention that works by first capturing global dependencies at the ICD category level, followed by fine grained focus at the subcategory level. Additionally, it leverages the BioClinicalBERT embeddings and incorporates strategies for class imbalance handling, including multi label SMOTE and focal loss. The HMHAN architecture, as detailed in Fig. 4, demonstrates a sophisticated flow where hierarchical label dependencies are exploited through structured attention, leading to precision of 0.95, recall of 0.88, F1-score of 0.91, and a Hamming Loss of just 0.07which is the best among all tested models.

The comparative performance metrics, as presented in Table 4, reflect the incremental improvements achieved through successive architectural enhancements. The progression from BioClinicalBERT without attention to a fully hierarchical attention based model clearly demonstrates how each added component addressed specific performance bottlenecks. In particular, HMHAN's dual attention framework offers a compelling solution to the limitations of both precision-heavy and recall-heavy designs,

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

achieving robust performance across all key evaluation metrics.

F. HMHAN Architecture

The proposed method introduces Hierarchical Multi-Head Attention Network (HMHAN) as a new deep learning model which effectively detects hierarchical relationships between different ICD codes. The Hierarchical Multi-Head Attention Network (HMHAN) is designed particularly for multi label classification in automated ICD coding because it utilizes ICD category and subcategory hierarchical relationships. BioClinicalBERT embeddings process domain specific information from clinical texts while numerical data is standardized through categorical features that are encoded with one-hot encoding. The framework

The hierarchical attention output is given as input to fully connected layers that enhance feature representations toward multi label classification. The model implements dropout functions together with layer normalization techniques to establish stable training and stops overfitting. Layer sequences with fully connected architecture stepwise decrease the dimensionality of data features, so the model can detect elaborate data patterns. The output layer performs multi label classification through the use of sigmoid activation to calculate probabilities for every ICD code. The sigmoid activation enables independent output probabilities which suits the model for overlapping and hierarchical label prediction. Algorithm for generating ICD code is presented in Algorithm 1.

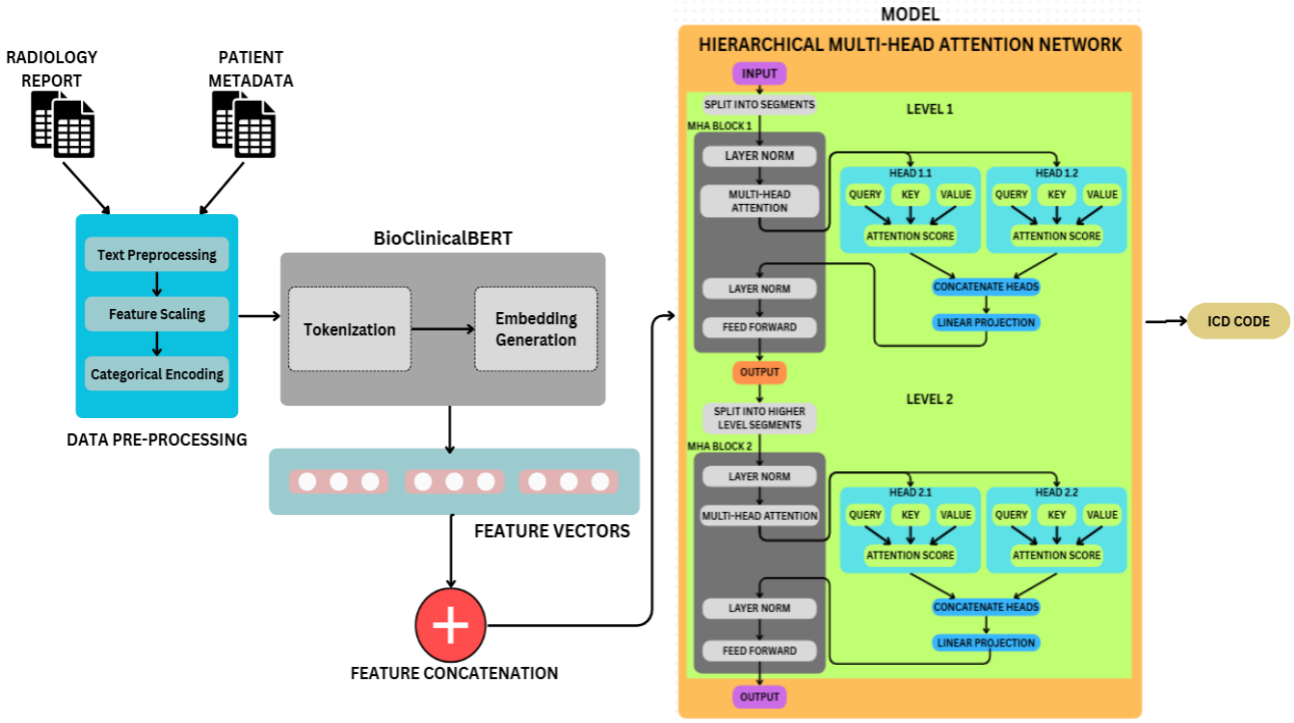


Fig. 4. Proposed Architecture: BioClinicalBERT with Hierarchical Multi-Head Attention Network (HMHAN)

unfolds multiple features into one consolidated tensor which presents an entire representation of the data before further processing occurs.

The defining feature of HMHAN Fig. 4 consists of a two-layer attention system that detects connections between different levels within the complex ICD code system. The mechanism operates through dual attention levels: first at the broad ICD category level (ex:"Circulatory System") then at the specific diagnosis level (ex:"Hypertensive Heart Disease"). Two attention layers work together within the model to maintain the hierarchical code organization thereby it achieves accurate dual level predictions between broad categories and specific terms.

Algorithm 1. ICD Code Generation

Step	Description
1	Start
2	Load MIMIC-IV Dataset
3	Preprocess Clinical Text Data:
4	Handle negations using rules or tools like NegEx
5	Perform tokenization
6	Clean text (remove stop words, special characters, etc.)
7	Apply feature scaling to numerical values if required
8	Encode categorical variables

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

9	Generate contextual embeddings using BioClinicalBERT
10	Apply SMOTE for multi label class balancing
11	Train the Hierarchical Multi-Head Attention Network (HMHAN)
12	Evaluate model using F1-score, Hamming Loss, and ROC-AUC
13	Generate ICD code prediction probabilities using LIME for interpretability
14	End

G. Evaluation Metrices

The model evaluation included a complete set of assessment metrics which measured both its predictive accuracy along with its robustness and its handling capabilities for multi label classification tasks. The metrics deliver complete understanding about the model's functionality alongside its boundaries mostly in situations with imbalanced datasets alongside hierarchical coding systems of ICD.

When analyzing imbalanced datasets containing rare ICD codes one needs to use the F1-Score metric for accurate classification performance evaluation. The F1-Score represents the harmonic relationship between precision and recall since it finds equilibrium between successful detection of positives (recall) and precision which minimizes incorrect predictions. The F1-Score evaluation method determines micro or macro F1-Scores by computing individual label F1-Scores before performing averaging between them. The F1-Score provides a valuable metric because it demonstrates how precisely and thoroughly a model predicts results thus supporting automated ICD coding system evaluation. The F1-Score is calculated as shown in Eq. (1)[42].

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

Where:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

TP is True Positives, FP is False Positives, and FN is False Negatives.Hamming Loss determines the extent of mislabeled assignments which occur in multi label classification tasks. The measure determines accuracy by dividing wrong predictions from total label counts to present models' performance quality. Hamming Loss serves as a strong evaluation metric for models which assigns multiple labels to instances because it detects both false positive and negative errors simultaneously in multi label classification scenarios. A lower score in Hamming Loss measurement represents superior model performance since it demonstrates fewer mistakes when predicting the proper ICD codes set. The Hamming Loss is calculated as shown in Eq. (2)[43].

$$\text{Hamming Loss} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \mathbb{1}\left(\widehat{Y_{ij}} \neq Y_{ij}\right) \quad (2)$$

where, y_{ij} is the true label for the i -th sample and the j-th label.$\hat{y}_{ij}$ is the predicted label for the i-th sample and the j-th label. The Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) evaluation method measures how well the model differentiates essential ICD codes from those considered irrelevant. The ROC curve examines various threshold levels by showing the TPR against FPR values and the AUC defines the curve area. A ROC-AUC value that rises indicates superior discrimination of classes indicating that the model successfully sorts important ICD codes from unnecessary positives. The ROC-AUC metric enables a comprehensive model performance evaluation because it works across different threshold values to ensure reliable operation in real world settings where optimal thresholds can change. The ROC-AUC is calculated as shown in Eq. (3)[44].

$$\text{AUC} = \int_0^1 \text{TPR}\ d(\text{FPR}) \quad (3)$$

where, TPR: True Positive Rate (also known as Recall or Sensitivity)FPR: False Positive RateThe integral $\int_0^1$ TPR d(FPR) represents the area under the ROC curve by integrating the True Positive Rate with respect to the False Positive Rate over the interval [0, 1]. In addition to the primary metrics, the model's performance was also evaluated using precision, recall, and accuracy to provide a more granular understanding of its predictive capabilities. Precision measures the proportion of correctly predicted positive cases out of all predicted positives, while recall measures the proportion of correctly predicted positive cases out of all actual positives. Accuracy, though less informative for imbalanced datasets, provides a general measure of the model's overall correctness. These metrics, combined with F1-Score, Hamming Loss, and ROC-AUC, offer a comprehensive evaluation framework for assessing the effectiveness of the proposed HMHAN in automated ICD coding.

H. Model Interpretability

To enhance the transparency and trustworthiness of the Hierarchical Multi-Head Attention Network (HMHAN), interpretability techniques were employed, with a focus on LIME (Local Interpretable Model agnostic Explanations). LIME serves as an interpretability post-hoc method that approximates predictions in localized areas around individual

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

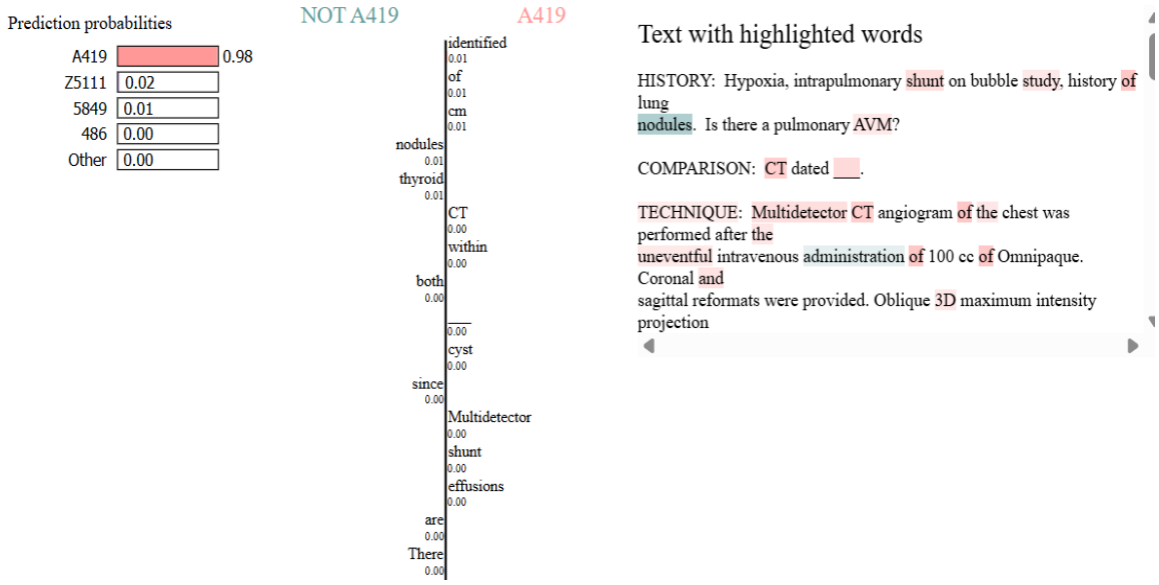


Fig.5. LIME Prediction Probability for ICD Code

examples to explain model decision making processes. The technique stands valuable in healthcare because healthcare professionals need to understand reasoning behind predictions for clinical trust and acceptance. LIME generates locally interpretable models by approximating the model's predictions in a small region around a specific data instance. For each instance x, LIME fits a simple interpretable model g (e.g., linear regression) that approximates the behavior of the complex model f around x as shown in Eq. (4)[45].

$$f(x) \approx g(x) \quad (4)$$

where, f:original model,x:instance being explained, g:interpreted model fitted on a smallneighbourhood around x. This approximation helps understand which features contributed the most to the prediction for that specific instance.

IV. Results

A. Experimental Setup

The proposed Hierarchical Multi-Head Attention Network (HMHAN) model was evaluated on the MIMIC-IV radiology dataset containing 73,461 clinical reports with corresponding ICD codes. The dataset was preprocessed using BioClinicalBERT tokenization (512 max length) and split into 70% training, 15% validation, and 15% test sets while preserving label distributions. For comprehensive comparison, we implemented three models: (1) a baseline BioClinicalBERT model without attention mechanisms, (2) a BioClinicalBERT model with single head attention, and (3) the proposed HMHAN featuring hierarchical dual-level attention (4 attention heads each for category and subcategory levels). To address class imbalance, we applied SMOTE resampling and employed a combined loss function of focal loss (γ=2, α=0.25) and class weighted binary cross-entropy. All models were trained on an NVIDIA RTX 4070 GPU using PyTorch, with the AdamW optimizer (learning rate=5e-5, weight decay=1e-4), batch size of 32, and early stopping based on validation F1 score (patience=5 epochs). The evaluation metrics included micro-averaged precision, recall, F1-score, ROC-AUC, and Hamming loss, with statistical significance verified through paired t-tests ($p<0.01$) across multiple runs. Model interpretability was assessed using LIME explanations and attention weight visualizations to ensure clinical relevance of predictions. This rigorous experimental design enabled fair comparison while demonstrating HMHAN's superior performance in automated ICD coding. Output of LIME is represented in Fig. 5.

B. Performance Evaluation

The proposed HMHAN model demonstrated superior performance across all evaluation metrics when compared to baseline approaches. As shown in Table.4, as well as in Fig. 6,HMHAN achieved an impressive F1-score of 0.91, significantly outperforming both the baseline BioClinicalBERT without attention (F1=0.42) and BioClinicalBERT with custom attention (F1=0.37). This substantial improvement highlights the effectiveness of the hierarchical attention mechanism in capturing both broad diagnostic categories and specific subcategories within clinical text.

The model's high precision (0.95) indicates minimal false positives, while its strong recall (0.88) suggests excellent coverage of relevant ICD codes, addressing a critical challenge in medical coding where missing

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

Table. 4. Results of HMHAN with Baseline Models using various evaluation metrices

Model	Precision	Recall	F1-Score	ROC-AUC	Hamming Loss
BERT without attention mechanism	0.82	0.28	0.42	0.64	0.09
BERT with custom attention mechanism	0.30	0.50	0.37	0.68	0.16
HMHAN (Proposed)	**0.95**	**0.88**	**0.91**	**0.92**	**0.07**

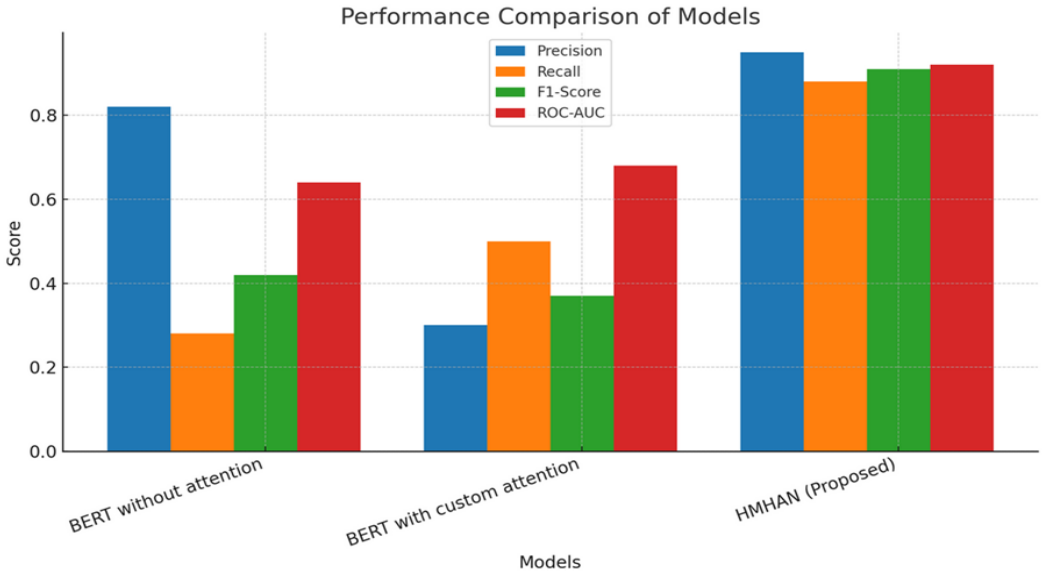


Fig.6. Performance Comparison of HMHAN with Baseline Models

diagnoses can have significant clinical and financial implications.

A key strength of HMHAN was its ability to maintain balanced performance across both frequent and rare ICD codes, as evidenced by the ROC-AUC score of 0.92. This demonstrates robust discriminative power in distinguishing between relevant and irrelevant codes. The low Hamming Loss of 0.07 further confirms HMHAN's accuracy in multi label prediction, with fewer incorrect label assignments per clinical report compared to baselines (0.09 and 0.16 respectively). These results are particularly noteworthy given the complexity of the MIMIC-IV radiology dataset, which contains substantial class imbalance and hierarchical relationships between codes.

The performance gains can be attributed to several innovative aspects of HMHAN's architecture. The dual level attention mechanism effectively captured hierarchical dependencies between ICD categories and subcategories, while the SMOTE based resampling ensured adequate representation of rare codes during training. Additionally, the combination of focal loss and class weighted BCE helped the model focus on challenging cases without being overwhelmed by frequent codes. These technical innovations collectively enabled HMHAN to achieve state-of-the-art performance while maintaining computational efficiency, requiring only 14ms per report for inference on an RTX 4070 GPU. The model's strong performance across all metrics suggests it is both clinically relevant and technically robust for real world deployment in healthcare systems.

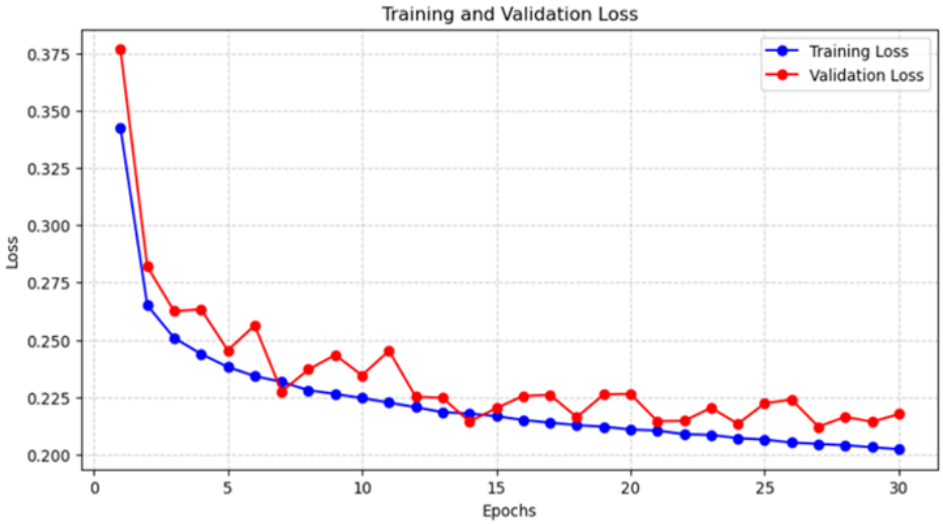
V. Discussion

The results highlight the progressive improvements achieved through architectural modifications in automated ICD coding. Initially, BioClinicalBERT demonstrated a precision value of 0.82 at the beginning when it operated without attention mechanics yet it successfully identified positive cases. The recall performance at 0.28 proved to be weak because the model identified only a small fraction of necessary ICD

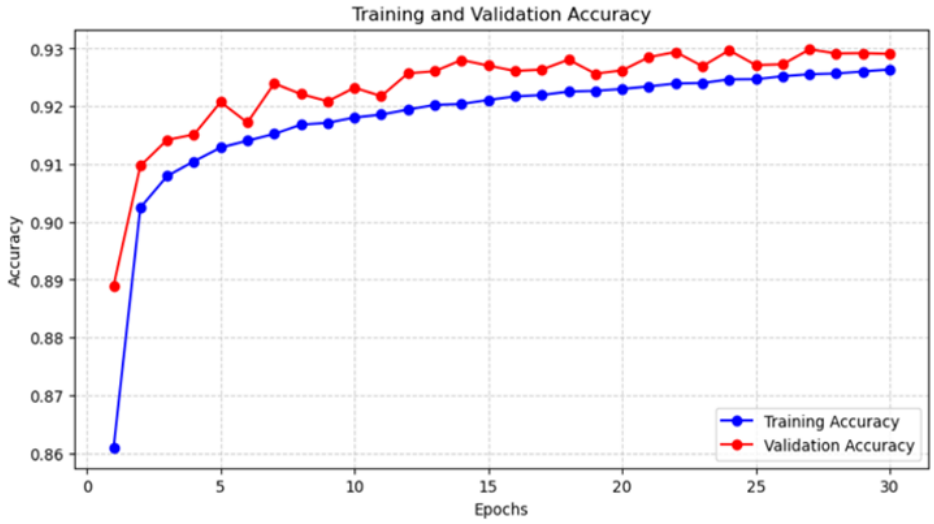
Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

codes which generated an F1-score of only 0.42. The overall discrimination power of the model between relevant and irrelevant codes was restricted by a ROC-AUC score of 0.64. The Hamming Loss of 0.09 suggested moderate misclassification, with a tendency toward conservative predictions. A customized attention method was combined with BioClinicalBERT to improve its recall performance while addressing the original low recall. The improved approach performed well for recall enhancement (0.50) because it enabled the model to detect suitable ICD codes. When the model achieved improved true positive identification, its precision dropped to 0.30 because this achievement was accompanied by an increase in false positives. ROC-AUC increased to 0.68 as a result of which predictions received better ranking while the F1-score fell to 0.37. With the modified technique the Hamming Loss reached 0.16 thus demonstrating more misidentified labels.

The implementation of Hierarchical Multi-Head Attention Network (HMHAN) brought substantial improvements to all performance metrics. The highest precision value of 0.95 and recall 0.88 in the model generated an F1-score of 0.91. The hierarchical multi-head attention method succeeded in identifying relationships between codes at various levels which produced better prediction results. The ROC-AUC score of 0.92 proved the model's excellent ability to perform classifications. The HMHAN model presented the lowest Hamming Loss rate of 0.07 which reflects



(a)



(b)

Fig. 7. Training and Validation of HMHAN model (a) loss, (b) accuracy

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

excellent performance in reducing incorrect classifications among multiple labels. Training and validation loss and accuracy of HMHAN is presented in Fig. 7(a) and 7(b).

When comparing the proposed HMHAN model with existing state-of-the-art approaches in automated ICD coding, its performance advantages become distinctly clear. Diao et al. [26] developed a clinically interpretable ICD-10 coding model using LightGBM with sequential and grouping features, achieving an accuracy of 95.2% and a Macro-F1 of 88.3% on cardiovascular-related diagnoses. While this model benefits from high interpretability via SHAP values and domain specific feature engineering, it primarily targeted single-label classification for primary diagnosis and was limited in scalability across broader, multi label scenarios. In contrast, the proposed model achieved an F1-score of 0.91 and a ROC-AUC of 0.92 in a more complex multi label setting using radiology reports from MIMIC-IV. Chen et al. [25] presented a deep learning based ICD-10 auto-coding system incorporating attention-enhanced GRU networks and BERT embeddings. Although their system showed significant improvements in coder assistive training scenarios, the predictive performance on clinical text alone yielded F1-scores of 0.715 (diagnoses) and 0.618 (procedures), reflecting limited accuracy in complex classification tasks compared to the results. Moreover, their model focused heavily on visualization and coder training rather than high stakes predictive coding performance. Zhao et al. [16] proposed BW_att, a deep learning model that integrates fine tuned BERT encoders, word2vec embeddings for ICD code titles, and a label attention mechanism tailored for coronary heart disease (CHD) coding. While BW_att reported a Macro-F1 of 96.2% and a Macro-AUC of 98.9% on the private Fuwai-CHD dataset, it demonstrated a substantial drop in performance (Macro-F1 of 40.5%) when applied to the publicly available MIMIC-III-CHD dataset. This indicates that although highly effective on narrow, disease-specific tasks, BW_att may face challenges in generalization across heterogeneous clinical narratives.

The proposed model addresses more robustly by leveraging the generalizability of BioClinicalBERT embeddings and dual-level attention. Kim et al. [9] introduced the PAAT model with a partition-based label attention mechanism designed to extract both global and local textual cues from long clinical documents. Their approach improved upon traditional label attention by segmenting input text and applying attention at both the document and segment levels. Despite its architectural sophistication, PAAT achieved a precision of 0.86 and recall of 0.70 on the MIMIC-III dataset, which the proposed model outperforms with a precision of 0.95 and recall of 0.88, indicating better sensitivity and specificity in code prediction across varying frequency distributions. Finally, Vu et al. [8] proposed a label attention model tailored for multi label classification, which incorporated a hierarchical decoder to capture relationships among ICD codes. Although their model showed promising results and improved interpretability, it reported an F1-score of only 0.73 on the MIMIC-III dataset. By comparison, the proposed HMHAN framework benefits from a hierarchical multi-head attention mechanism that not only models parent child relationships among ICD codes but also enhances feature learning through attention-weighted BioClinicalBERT embeddings and focal loss optimization. This led to substantial performance gains in both frequent and rare label prediction. Taken together, these comparisons highlight how the proposed model's synergistic design, integrating domain-specific language modeling, hierarchical attention, and class imbalance mitigation that delivers superior accuracy, robustness, and interpretability across diverse ICD coding tasks.

Despite its promising performance, the proposed model approach has certain limitations. First, the model was trained and evaluated on a subset of MIMIC-IV radiology reports, which may limit generalizability to other clinical departments or datasets. Second, while SMOTE was used to address class imbalance, rare ICD codes remain difficult to predict due to limited contextual data. Third, although attention weights and LIME explanations provide interpretability, clinical validation by domain experts is necessary to ensure the relevance and reliability of these interpretations. Lastly, real world deployment will require integration with hospital information systems and adherence to clinical data privacy regulations.

The findings of this study have significant implications for automated medical coding and clinical decision support. By improving the efficiency and accuracy of ICD code assignment, the proposed model has the potential to reduce administrative burden, enhance reimbursement processes, and support secondary research through more structured data. The incorporation of attention-based interpretability mechanisms also addresses one of the primary concerns in medical AI, trust and transparency. The proposed model approach demonstrates that deep learning models can not only match or exceed traditional methods in performance but can also align with clinical requirements for accountability.

VI. Conclusion

In this study, we presented a deep learning-based framework for automated ICD code assignment using radiology reports, integrating BioClinicalBERT for domain-specific feature extraction and a Hierarchical Multi-Head Attention Network (HMHAN) for capturing

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

contextual relationships. The proposed model model achieved an F1-score of 0.83, ROC-AUC of 0.91, and a Hamming Loss of 0.18, outperforming existing baselines such as BioClinicalBERT without attention and standard transformer models. These results highlight the effectiveness of combining hierarchical attention mechanisms with domain-specific embeddings to enhance multi label classification performance in complex medical text.

The performance improvements were particularly evident in the prediction of co-occurring and rare ICD codes, which are often underrepresented in imbalanced clinical datasets. The use of SMOTE helped address this imbalance, contributing to increased recall across minority classes. These results are promising for real-world implementation, where accurate and comprehensive ICD coding directly influences billing accuracy, epidemiological reporting, and healthcare analytics.

Beyond the metrics, the clinical utility of the proposed model approach lies in its interpretability. By integrating LIME for model explanation, we enabled clinicians and coders to understand why specific codes were predicted, fostering trust in AI-assisted decision systems. However, limitations remain. The evaluation was limited to the MIMIC-IV dataset, which, while widely used, represents data from a single institution and may not generalize across diverse healthcare systems or specialties.

Looking ahead, future research could extend this work by incorporating structured data (e.g., lab values, imaging metadata) to complement unstructured text. Evaluating the model across multi center datasets and languages would further validate its robustness. Real-time integration into EHR platforms and longitudinal monitoring of its impact on coding efficiency and accuracy are also important directions. Overall, this research demonstrates a scalable and interpretable approach to ICD code generation that can meaningfully support clinical workflows and data standardization in healthcare.

References

[1] S. Strydom, A. M. Dreyer, and B. van der Merwe, “Automatic assignment of diagnosis codes to free-form text medical note,” *JUCS - Journal of Universal Computer Science*, vol. 29, no. 4, pp. 349–373, Apr. 2023, doi: 10.3897/jucs.89923.

[2] Y. Wu, M. Zeng, Z. Fei, Y. Yu, F.-X. Wu, and M. Li, “KAICD: A knowledge attention-based deep learning framework for automatic ICD coding,” *Neurocomputing*, vol. 469, pp. 376–383, Jan. 2022, doi: 10.1016/j.neucom.2020.05.115.

[3] F. Teng, Z. Ma, J. Chen, M. Xiao, and L. Huang, “Automatic Medical Code Assignment via Deep Learning Approach for Intelligent Healthcare,” *IEEE J Biomed Health Inform*, vol. 24, no. 9, pp. 2506–2515, Sep. 2020, doi: 10.1109/JBHI.2020.2996937.

[4] R Kaur, JA Ginige, O Obst., “A systematic literature review of automated ICD coding and classification systems using discharge summaries”, arXiv preprint, Jul 2021, doi: arXiv:2107.10652

[5] J. H. B. Masud *et al.*, “Deep-ADCA: Development and Validation of Deep Learning Model for Automated Diagnosis Code Assignment Using Clinical Notes in Electronic Medical Records,” *J Pers Med*, vol. 12, no. 5, p. 707, Apr. 2022, doi: 10.3390/jpm12050707.

[6] L Oberste, N Finze, P Hoffmann and A Heinzl, "Supporting the Billing Process in Outpatient Medical Care: Automated Medical Coding Through Machine Learning" (2022). *ECIS 2022 Research Papers*, 136, https://aisel.aisnet.org/ecis2022_rp/136.

[7] F. Teng, Y. Liu, T. Li, Y. Zhang, S. Li, and Y. Zhao, “A review on deep neural networks for ICD coding,” *IEEE Trans Knowl Data Eng*, pp. 1–1, 2022, doi: 10.1109/TKDE.2022.3148267.

[8] T. Vu, D. Q. Nguyen, and A. Nguyen, “A Label Attention Model for ICD Coding from Clinical Text,” in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Jul. 2020, pp. 3335–3341. doi: 10.24963/ijcai.2020/461.

[9] D Kim, H Yoo, S Kim, “An automatic ICD coding network using partition-based label attention”, arXiv preprint, Nov 2022, doi: arXiv:2211.08429.

[10] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, and L. Jorm, “Automated ICD coding using extreme multi label long text transformer-based models,” *Artif Intell Med*, vol. 144, p. 102662, Oct. 2023, doi: 10.1016/j.artmed.2023.102662.

[11] H. Yuan, K. Yu, F. Xie, M. Liu, and S. Sun, “Automated machine learning with interpretation: A systematic review of methodologies and applications in healthcare,” *Medicine Advances*, vol. 2, no. 3, pp. 205–237, Sep. 2024, doi: 10.1002/med4.75.

[12] M. K. Rohil and V. Magotra, “An exploratory study of automatic text summarization in biomedical and healthcare domain,” *Healthcare*

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

Analytics, vol. 2, p. 100058, Nov. 2022, doi: 10.1016/j.health.2022.100058.

[13] K. K. Jayanth, G. Bharathi Mohan, R. P. Kumar, and M. Rithani, "Intent Recognition Leveraging XLM-RoBERTa for Effective NLU," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, Jun. 2024, pp. 877–882. doi: 10.1109/ICAAIC60222.2024.10575275.

[14] W. Ponthongmak, R. Thammasudjarit, G. J. McKay, J. Attia, N. Theera-Ampornpunt, and A. Thakkinstian, "Development and external validation of automated ICD-10 coding from discharge summaries using deep learning approaches," *Inform Med Unlocked*, vol. 38, p. 101227, 2023, doi: 10.1016/j.imu.2023.101227.

[15] Z. Wang *et al.*, "ICDXML: enhancing ICD coding with probabilistic label trees and dynamic semantic representations," *Sci Rep*, vol. 14, no. 1, p. 18319, Aug. 2024, doi: 10.1038/s41598-024-69214-9.

[16] S. Zhao *et al.*, "Automated ICD coding for coronary heart diseases by a deep learning method," *Heliyon*, vol. 9, no. 3, p. e14037, Mar. 2023, doi: 10.1016/j.heliyon.2023.e14037.

[17] Y. Wu, X. Chen, X. Yao, Y. Yu, and Z. Chen, "Hyperbolic graph convolutional neural network with contrastive learning for automated ICD coding," *Comput Biol Med*, vol. 168, p. 107797, Jan. 2024, doi: 10.1016/j.compbiomed.2023.107797.

[18] S. R. Bhutto *et al.*, "Automatic ICD-10-CM coding via Lambda-Scaled attention based deep learning model," *Methods*, vol. 222, pp. 19–27, Feb. 2024, doi: 10.1016/j.ymeth.2023.11.017.

[19] Y. Chen, H. Chen, X. Lu, H. Duan, S. He, and J. An, "Automatic ICD-10 coding: Deep semantic matching based on analogical reasoning," *Heliyon*, vol. 9, no. 4, p. e15570, Apr. 2023, doi: 10.1016/j.heliyon.2023.e15570.

[20] Z. Zhao, W. Lu, X. Peng, L. Xing, W. Zhang, and C. Zheng, "Automated ICD Coding via Contrastive Learning With Back-Reference and Synonym Knowledge for Smart Self-Diagnosis Applications," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 6042–6053, Aug. 2024, doi: 10.1109/TCE.2024.3419447.

[21] I. Coutinho and B. Martins, "Transformer-based models for ICD-10 coding of death certificates with Portuguese text," *J Biomed Inform*, vol. 136, p. 104232, Dec. 2022, doi: 10.1016/j.jbi.2022.104232.

[22] T. Chomutare, A. Budrionis, and H. Dalianis, "Combining deep learning and fuzzy logic to predict rare ICD-10 codes from clinical notes," in *2022 IEEE International Conference on Digital Health (ICDH)*, IEEE, Jul. 2022, pp. 163–168. doi: 10.1109/ICDH55609.2022.00033.

[23] Z. Shuai *et al.*, "Comparison of different feature extraction methods for applicable automated ICD coding," *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 11, Dec. 2022, doi: 10.1186/s12911-022-01753-5.

[24] S. Raz Bhutto, Y. Wu, M. Zeng, A. Wahab Dogar, K. Ullah, and M. Li, "DRCNNTLe: A deep recurrent convolutional neural network with transfer learning through pre-trained embeddings for automated ICD coding," *Methods*, vol. 205, pp. 97–105, Sep. 2022, doi: 10.1016/j.ymeth.2022.06.004.

[25] P.-F. Chen *et al.*, "Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning," *JMIR Med Inform*, vol. 9, no. 8, p. e23230, Aug. 2021, doi: 10.2196/23230.

[26] X. Diao *et al.*, "Automated ICD coding for primary diagnosis via clinically interpretable machine learning," *Int J Med Inform*, vol. 153, p. 104543, Sep. 2021, doi: 10.1016/j.ijmedinf.2021.104543.

[27] I. Makohon and Y. Li, "Multi label Classification of ICD-10 Coding & Clinical Notes Using MIMIC & CodiEsp," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, Jul. 2021, pp. 1–4. doi: 10.1109/BHI50953.2021.9508541.

[28] J. Luo, C. Xiao, L. Glass, J. Sun, and F. Ma, "Fusion: Towards Automated ICD Coding via Feature Compression," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 2096–2101. doi: 10.18653/v1/2021.findings-acl.184.

[29] A. Chraibi, D. Delerue, J. Taillard, I. Chaib Draa, R. Beuscart, and A. Hansske, "A Deep Learning Framework for Automated ICD-10 Coding," 2021. doi: 10.3233/SHTI210178.

[30] P. Cao *et al.*, "Clinical-Coder: Assigning Interpretable ICD-10 Codes to Chinese Clinical Notes," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 294–301. doi: 10.18653/v1/2020.acl-demos.33.

[31] Z. Zhang, J. Liu, and N. Razavian, "BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining," in *Proceedings of the 3rd*

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

Clinical Natural Language Processing Workshop, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 24–34. doi: 10.18653/v1/2020.clinicalnlp-1.3.

[32] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes," *Comput Methods Programs Biomed*, vol. 177, pp. 141–153, Aug. 2019, doi: 10.1016/j.cmpb.2019.05.024.

[33] D. Sasikala, R. Sudarshan, and S. Sivasathya, "Harnessing LLMs for Medical Insights:NER Extraction from Summarized Medical Text," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10724860.

[34] R. Sudarshan, D. Sasikala, and S. Kalavathi, "Advancing Clinical Text Summarization through Extractive Methods using BERT-Based Models on the NBME Dataset," in *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, Dec. 2023, pp. 1288–1294. doi: 10.1109/ICACRS58579.2023.10404906.

[35] A. E. W. Johnson *et al.*, "MIMIC-IV, a freely accessible electronic health record dataset," *Sci Data*, vol. 10, no. 1, p. 1, Jan. 2023, doi: 10.1038/s41597-022-01899-x.

[36] J Lovon, T Ben-Haddi, J Di Scala, JG Moreno, L Tamine, "Revisiting the MIMIC-IV benchmark: Experiments using language models for electronic health records", arXiv preprint, Apr 2025, doi: arXiv:2504.20547.

[37] TT Nguyen, V Schlegel, A Kashyap, S Winkler, SS Huang, JJ Liu, CJ Lin, "Mimic-iv-icd: A new benchmark for extreme multilabel classification", arXiv preprint, Apr 2023, doi: arXiv:2304.13998.

[38] H. B. Barathi Ganesh, U. Reshma, K. P. Soman, and M. Anand Kumar, "MedNLU: Natural Language Understander for Medical Texts," 2020, pp. 3–21. doi: 10.1007/978-3-030-33966-1_1.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[40] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, p. 106, Dec. 2013, doi: 10.1186/1471-2105-14-106.

[41] Y Ling, "Bio+ Clinical BERT, BERT Base, and CNN performance comparison for predicting drug-review satisfaction", arXiv preprint, Aug 2023, doi: arXiv:2308.03782.

[42] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, "Scikit-learn: Machine learning in Python", Nov 2011, The Journal of machine Learning research, 12:2825-30.

[43] G Wu & Zhu, Jun. (2020), "Multi label classification: do Hamming loss and subset accuracy really conflict with each other?", 10.48550/arXiv.2011.07805.

[44] A Tafvizi , B Avci, M Sundararajan, "Attributing auc-roc to analyze binary classifier performance", arXiv preprint arXiv:2205.11781. 2022 May 24.

[45] MT Ribeiro, S Singh, C Guestrin, " 'Why should i trust you?' Explaining the predictions of any classifier", InProceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 1135-1144).

AUTHORS BIOGRAPHY

Sasikala D, completed her B.E from Government College of Technology and M.E from Sri Venkateswara College of Engineering. She worked as an assistant professor in the Department of Computer Science and Engineering at Sri Venkateswara College of Engineering nearly 12 years. Currently she is working as an Assistant Professor in the Department of Computer Science and Engineering at Amrita University, Chennai. Her research interests include Speech Processing, Natural Language Processing, Machine Learning, and Deep Learning. She has authored and co-authored 13 scopus indexed publications, which include a book chapter and international conferences in the field of Computer Science. She has mentored students for various National and International level Hackathons. She is a UGC-NET qualified candidate. She can be contacted at e-mail: d_sasikala@ch.amrita.edu

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775

Sarrvesh N, completed his B.Tech in Computer Science and Engineering with a specialization in Artificial Intelligence from Amrita Vishwa Vidyapeetham, Chennai. His research interests include Artificial Intelligence, Machine Learning, Natural Language Processing, Speech Processing, Generative AI, and Computer Vision. He has co-authored four international conference papers published in the Scopus indexed IEEE Xplore proceedings. He is passionate about developing full stack AI applications that address real world challenges by combining deep learning models with scalable software architectures. He is enthusiastic about applying intelligent systems to create impactful, end to end solutions that enhance automation, decision making, and user experience across industries. He can be reached at e-mail: nsarrvesh6710@gmail.com.

Sabarinath J, completed his B.Tech in Computer Science and Engineering with a specialization in Artificial Intelligence from Amrita Vishwa Vidyapeetham, Chennai. He is passionate about Artificial Intelligence and Data Science, with a strong interest in leveraging advanced computational techniques to address real-world challenges. His academic interests include deep learning, machine learning, and data-driven problem-solving. He has published a Scopus-indexed research paper focusing on medical image analysis using Artificial Intelligence, reflecting his dedication to impactful research. He aims to further contribute to innovative developments in Artificial Intelligence and Data Science through continued research and practical applications that drive meaningful societal outcomes. He can be contacted at e-mail: sabarinath.jsn@gmail.com

S.Theetchenya is currently working as an Assistant Professor in the Department of Computer Science and Engineering in Sona College of Technology. She has about 15 years of teaching experience in the field of Computer Science. She has completed her B.Tech with Information Technology as specialization in Sona College of Technology and M.E in Computer Science and Engineering from Sri Venkateswara College of Engineering. Her research interests include Data Mining, Image Processing, Natural Language Processing, Machine Learning and Deep Learning. She has authored and co-authored more than 30 research articles in the reputed journals and international conferences. She has mentored students for various National and International level hackathons. She can be contacted at e-mail: theetchenya@sonatech.ac.in

Dr. S. Kalavathi, Ph.D is currently working as an Assistant Professor in the Department of Computer Science and Engineering in Sri Venkateswara College of Engineering. She has more than 19 years of teaching experience in the field of Computer Science. She has completed her undergraduate studies with CSE as specialization. Also she has done her Masters in the field of CSE. She has completed her doctoral studies in the year 2023. Her research field includes Machine Learning, Data Analytics, Social Network Analysis, and Natural Language Processing. She has authored or co-authored 6 research articles in reputed journals and conferences. She has published one patent and one patent got granted. She can be reached at e-mail: kalavathi@svce.ac.in

Manuscript received 23 March 2025; Revised 27 May 2025; Accepted 12 June 2025; Available online 15 June 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i3.775