

AI-Enhanced Lung Cancer Prediction: A Hybrid Model's Precision Triumph

Cyrille Yetuyetu Kesiku  and Begonya Garcia-Zapirain 

Abstract—Lung cancer is considered one of the most dangerous cancers, with a 5-year survival rate, ranking the disease among the top three deadliest cancers globally. Effectively combating lung cancer requires early detection for timely targeted interventions. However, ensuring early detection poses a major challenge, giving rise to innovative approaches. The emergence of artificial intelligence offers revolutionary solutions for predicting cancer. While marking a significant healthcare shift, the imperative to enhance artificial intelligence models remains a focus, particularly in precision medicine. This study introduces a hybrid deep learning model, incorporating Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory Networks (BiLSTM), designed for lung cancer detection from patients' medical notes. Comparative analysis with the MIMIC IV dataset reveals the model's superiority, achieving an MCC of 96.2% with an Accuracy of 98.1%, and outperforming LSTM and BioBERT with an MCC of 93.5%, an accuracy of 97.0% and MCC of 95.5 with an accuracy of 98.0% respectively. Another comprehensive comparison was conducted with state-of-the-art results using the Yelp Review Polarity dataset. Remarkably, our model significantly outperforms the compared models, showcasing its superior performance and potential impact in the field. This research signifies a significant stride toward precise and early lung cancer detection, emphasizing the ongoing necessity for Artificial Intelligence model refinement in precision medicine.

Index Terms—Lung cancer, text classification, CNN BiLSTM attention, natural language processing, deep learning model.

I. INTRODUCTION

DESPITE advancements in medical technology and Diagnostic strategies, lung cancer remains one of the most formidable challenges in modern healthcare, characterized by its high mortality rates and the complexity of its diagnosis and treatment. With a 5-year [1] survival rate placing it among the top three deadliest cancers globally, the urgency to develop effective strategies for early detection and intervention is paramount. However, conventional diagnostic methods often fail to facilitate timely detection, highlighting the critical need for innovative

approaches to enhance early diagnosis and improve patient outcomes. The emergence of artificial intelligence (AI) has ushered in a new era of healthcare innovation, offering promising solutions for predictive modeling and disease detection. In particular, the application of deep learning techniques in medical image analysis [2] and clinical data interpretation has shown significant potential for revolutionizing cancer diagnosis and treatment [3]. Despite these advancements, there exists a gap [4] in the literature regarding the development of robust Artificial Intelligence models specifically tailored for the early detection of lung cancer from patients' medical records.

To address this gap, this study proposes a novel hybrid deep learning model integrating Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory Networks (BiLSTM) for the detection of lung cancer from patients' medical notes. By leveraging the rich textual information contained within medical records, the proposed model aims to improve the accuracy and efficiency of lung cancer diagnosis, ultimately leading to earlier detection and more targeted interventions. This research endeavors to make a significant stride toward precise and early detection of lung cancer, underscoring the ongoing necessity for Artificial Intelligence model refinement and innovation in the field of precision medicine.

Building upon the foundation laid by traditional machine learning models, recent advancements in deep learning, particularly in the realm of natural language processing (NLP), have revolutionized text classification [9] in healthcare. Analyzing historical textual data to discern social [7] habits and understand family [8] and medical histories can be a daunting task. However, the utilization of NLP techniques streamlines this effort by extracting pertinent and contextual information from patients' textual documents. This invaluable assistance to healthcare professionals enhances the medical decision-making process, thereby improving patient outcomes. Traditionally, machine learning models such as Support Vector Machines (SVM) [11], Decision Trees [12], Naive Bayes [10], and ensemble learning [13] approaches have yielded significant advancements in text retrieval. However, the emergence of deep learning models has surpassed these traditional methods, particularly in the domain of biomedical text classification. Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), have demonstrated promising performance in processing sequential [14] data, thereby enhancing contextual understanding in biomedical text classification tasks. BiLSTM [24], in particular, leverages its unique architecture to capture both past and future contextual information,

Received 9 February 2024; revised 15 April 2024 and 24 July 2024; accepted 14 August 2024. Date of publication 22 August 2024; date of current version 9 September 2025. (Corresponding author: Cyrille Yetuyetu Kesiku.)

The authors are with the Engineering Faculty, Department of Computer Science, Electronics and Telecommunications, University of Deusto, 48007 Bilbao, Spain (e-mail: cyrille.kesiku@opendeusto.es; mbgarciazapi@deusto.es).

Digital Object Identifier 10.1109/JBHI.2024.3447583

addressing the challenge of gradient vanishing and improving overall performance.

Moreover, Convolutional Neural Networks (CNNs) have transformed text classification, especially in the biomedical domain, due to their ability to extract local features efficiently. One-dimensional CNNs (1D CNNs) [23] are particularly well-suited for sequence classification tasks, including text data, as they employ one-dimensional filters to extract relevant features from sequences. This convergence of deep learning architectures and NLP techniques has paved the way for more sophisticated approaches to text classification, with significant implications for precision medicine and healthcare decision-making.

In this study, we aim to build upon these advancements by proposing a hybrid deep-learning model that integrates CNN and BiLSTM architectures for the early detection of lung cancer from patients' medical notes. By leveraging the capabilities of both CNN and BiLSTM in capturing local and long-range dependencies, respectively, our model seeks to enhance the accuracy and efficiency of lung cancer diagnosis. Through comparative analysis with external datasets, we endeavor to demonstrate the superiority of our proposed model and its potential to contribute to precision medicine approaches in oncology.

The convolution equation in one-dimensional convolutional neural networks (1D CNN) is formally stated as follows:

$$x_i = \sigma(W \cdot [e_{i:i+k-1}] + b) \quad (1)$$

Here $e_{i:i+k-1}$ is used to denote the sub-sequence of word embeddings starting from location i and ending at position $i + k - 1$. The filter weight matrix, denoted as W , is a key component in the context of this discussion. And b represents the bias. Other methods that constitute the state of the art in the field of NLP are transformers; these algorithms, which are based on attention mechanisms, are highly effective in text classification tasks, and this attention in a transformer is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \quad (2)$$

Where Q , K , and V represent the query, key, and value matrices respectively, and d_k refers to the key dimension.

In recent years, research in the field of text analysis has seen significant growth, with a particular interest in comparing and combining different neural network models to improve text prediction accuracy. Let us examine various studies that explore CNN models versus RNNs and their fusion to enhance text prediction accuracy.

Yin et al. (2017) [53] conducted pioneering work comparing CNNs and RNNs across multiple natural language processing (NLP) tasks, showcasing the superior performance of the GRU model in text classification. Their work laid the groundwork for the effective use of RNNs in text classification tasks. Soni et al. (2023) [54] introduced TextConvoNet, an innovative CNN architecture extracting both intra-sentence and inter-sentence features, surpassing existing models in text classification. This study highlighted the importance of multi-level feature extraction to improve the performance of text classification models. Li Xiaoyan et al. (2022) [55] proposed a GloVe-CNN-BiLSTM

model amalgamating diverse architectures for sentiment analysis, exhibiting exceptional performance and scalability. Their approach demonstrated the effectiveness of integrating different neural network architectures to enhance model accuracy in complex tasks such as sentiment analysis.

Jang et al. (2020) [56] devised a hybrid Bi-LSTM + CNN model with attention, outperforming individual models in terms of precision, recall, and F1 score, particularly in movie review classification. Their research underscored the importance of integrating attention mechanisms to capture relevant information in textual data.

In comparison with previous studies, our research has achieved outstanding performance in lung cancer classification. By leveraging a novel approach that integrates 1D CNN, BiLSTM, and Attention architectures in parallel, our model has demonstrated superior results. These findings underscore the effectiveness of our approach in handling complex medical data, offering new insights for early detection of this disease.

Extracting medical concepts from patients' historical medical records is a pressing issue in natural language processing, holding significant importance for various reasons. It facilitates the extraction of relevant information pertaining to treatment, diagnosis, and research, encapsulating specific behaviors and symptoms indicative of various diseases. This extracted information plays a crucial role in seeking specific answers to biomedical objectives, contributing to advancements in medical research and patient care. Building upon previous research endeavors, the subsequent studies have delved into addressing the challenge of medical concept extraction using diverse natural language processing and deep learning algorithms, notably leveraging CNN, RNN, and Transformer architectures.

In their study, MITHUN et al. (2023) [14] tackled the task of concept extraction specifically for lung cancer patients. Given the high prevalence and mortality rate associated with lung cancer, the authors developed a concept extraction pipeline tailored to this condition. They employed the BiLSTM model to classify extracted concepts, achieving promising performance outcomes. Another notable study by GEHRMANN et al. (2018)

Moreover, YANG et al. (2023) [17] conducted a comparative analysis of the BioBERT algorithm and conventional machine learning algorithms for classifying medical radiography data. Their findings highlighted the superior performance of the BioBERT algorithm, underscoring the potential of advanced transformer-based models in biomedical data classification tasks. In addition to these studies, other research endeavors have addressed various biomedical text extraction and classification challenges, utilizing a range of techniques including LSTM, R-CNN, and BERT [18], [19], [20]. Collectively, these studies contribute to advancing the field of medical text analysis and hold promise for enhancing healthcare outcomes through improved data-driven decision-making.

Classification of medical texts, particularly those related to lung cancer, requires a fine-tuned understanding of specialist terms, contextual nuances and semantic relationships. Our approach strives to overcome these complex challenges by exploiting a deep architecture that judiciously combines the power of the 1D Convolutional Neural Network (CNN) model with the

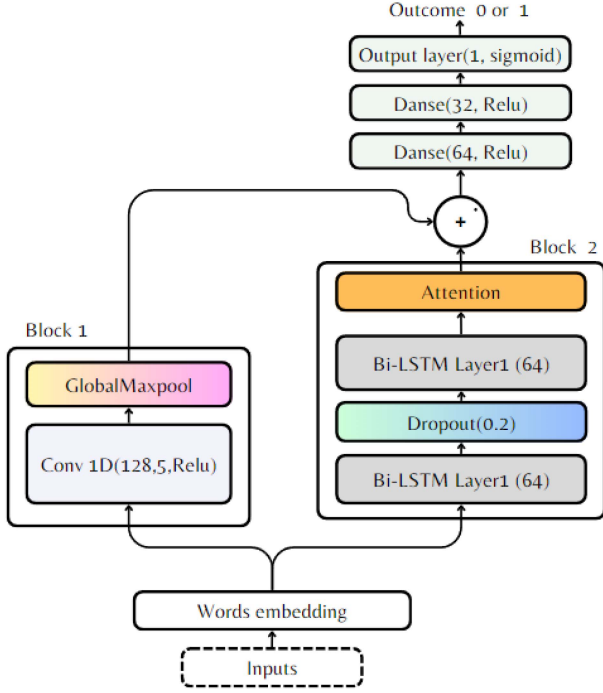


Fig. 1. CNN-BiLSTM_Attention Architecture. Initially, the text sequence undergoes an embedding operation. Two separate blocks then operate in parallel, feeding off the embedded vector. The first block, consisting of a 1D CNN followed by a global Max Pooling layer, focuses on the local extraction of text features. On the other hand, the second block implements a two-layer BiLSTM architecture, accompanied by a DropOut layer after the first BiLSTM layer and an Attention layer at the output of the second BiLSTM layer. The results obtained from these two blocks are then concatenated to feed the subsequent dense layers, thus completing the binary classification task.

dynamic contextualisation of the attention-enabled Bidirectional Long Short-Term Memory (BiLSTM). Our major contributions to this research can be summarised as follows:

- The research proposes a new deep learning model that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory Networks (BiLSTMs) with an attention mechanism for lung cancer detection from patients' medical notes.
- The model achieves state-of-the-art performance With a relatively low number of trainable parameters (12.5 million), on the MIMIC IV dataset, outperforming established benchmarks like BioBERT and LSTMs in terms of accuracy, AUC, and MCC.
- The model demonstrates strong generalizability beyond the medical domain by achieving excellent performance on the Yelp review polarity dataset, a benchmark for text classification tasks.

II. PROPOSED MODEL ARCHITECTURE

This study introduces an innovative NLP model that is specifically tailored to effectively address the challenging task of text classification within the medical domain. Fig. 1 shows the global architecture proposed in this study. The architecture begins with an input layer, followed by an embedding layer, which serves as

the foundation for the subsequent processing stages. Central to our architecture are two parallel blocks, each serving a distinct yet complementary purpose. The first block is comprised of a 1D Convolutional Neural Network (CNN) layer followed by a Global Max Pooling layer. This combination is adept at extracting local features from sequential data, such as word sequences in NLP tasks. Meanwhile, the second block is composed of two Bidirectional Long Short-Term Memory (BiLSTM) layers with dropout regularization applied between them, culminating in an attention layer. This configuration enables the model to capture both short-term and long-term dependencies within the input sequences, effectively enhancing its ability to understand context. Both blocks leverage the embeddings provided by the embedding layer, allowing them to extract rich semantic representations from the input data. The outputs of these blocks are then concatenated to create a fused representation, which is subsequently fed into a series of Dense layers for further processing. These Dense layers, consisting of three tiers, progressively distill and refine the information extracted from the input sequences, ultimately yielding the final prediction output. In the subsequent sections, we delve into a detailed exposition of each layer within our architecture, elucidating their roles and contributions towards the overall efficacy of the model in tackling NLP tasks.

A. Input Layer

The model processes a document $x_i \in X$, where x_i encompasses medical history, social history, familial history, and the patient's present illness.

Let X be our training dataset, made up of N text documents. Each document x_i is a sequence of words $w_{i,1}, w_{i,2}, \dots, w_{i,T_i}$, where T_i is the length of the document sequence i .

$$X = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \quad (3)$$

Here, (y_i) represents the class label 0 or 1 respectively no lung cancerous and lung cancerous for the document (x_i) . The aim is to learn a dense vector representation for each word in the corpus so that these embeddings may capture the semantic and contextual relationships in natural language.

B. Embedding Layer

We utilize in our architecture the Skip-gram [22] model as an embedding model within the TensorFlow.Keras library. This embedding model captures semantic and syntactic associations among words. It learns distributed vector representations for individual words, wherein similar words are represented by comparable vectors in a vector space. Moreover, through the utilization of the context window, the model adeptly grasps the contextual nuances of word occurrences, facilitating a deeper comprehension of meaning and syntax. Additionally, it is good for handling less common words by generating representations for such words based on their contextual usage within the corpus.

The Word2Vec model, as proposed by Mikolov et al. (2013) [21], operates on the principle that words with similar meanings possess comparable vector representations. The algorithm provides two primary methodologies, namely Skip-Gram

and Continuous Bag of Words (CBOW) [21]. Let us consider V as the vocabulary of our biomedical corpus, N as the embedding dimension, and D as the context dimension. Formalization of the likelihood that a word w_t is contextually close to another word w_c within a certain frame may be expressed in the Skip-Gram model as follows: The conditional probability of a context word w_c given a target word w_t may be calculated using the following formula:

$$P(w_c|w_t) = \frac{\exp(v'_{w_c} \cdot v_{w_t})}{\sum_{i=1}^{|V|} \exp(v'_i \cdot v_{w_t})} \quad (4)$$

Where v'_{w_c} represents the vector representation of the context word w_c , v . Word embeddings w_t and w_c are denoted as v_{w_t} and v'_{w_c} correspondingly, and these embeddings are acquired via a learning process to optimize the likelihood of contextual terms within the corpus.

Within the context of our study, the binary classification challenge involves representing each word in a biological document using its corresponding embedding. The vector representation is then used as input for our model, which consists of two parallel blocks.

C. Block 1: 1-Dimension CNN Layer and Global Max Pool

The first block of our architecture consists of a one-dimensional convolutional layer (1D-CNN) [23] followed by a global max-pooling operation. This configuration, characterized by the parameters (128,5,Relu), involves using 128 filters of size 5 with a ReLU activation function. In this bloc, the convolutional layer extracts local features from sequential data. By applying multiple filters to different parts of the word sequence, the model learns rich and informative representations that capture important patterns in the text. The use of the ReLU activation function after the convolution introduces non-linearity into the model, allowing it to better capture the complex relationships between words and improve its ability to model sequential data. The global max-pooling operation after the convolutional layer reduces the dimensionality of the extracted representations while preserving the most significant features. This helps reduce the number of parameters in the model, thus avoiding overfitting, while still maintaining a dense and informative representation of the input data.

Let x be a vector representing a sequence of words in a biomedical document, and W a filter weight matrix. The convolution required to obtain a new feature c_i from a window $x_{i:i+k-1}$ of words is defined as follows:

$$c_i = f(W \cdot x_{i:i+k-1} + b) \quad (5)$$

where f refers to an activation function, b to the bias, and k to the filter window size. Global Maxpooling [23] extracts the most important feature from the set of features extracted by convolution, guaranteeing a concise representation of the sequence's most salient features. This is formulated as follows:

$$\text{Global Maxpooling}(C) = \max(C) \quad (6)$$

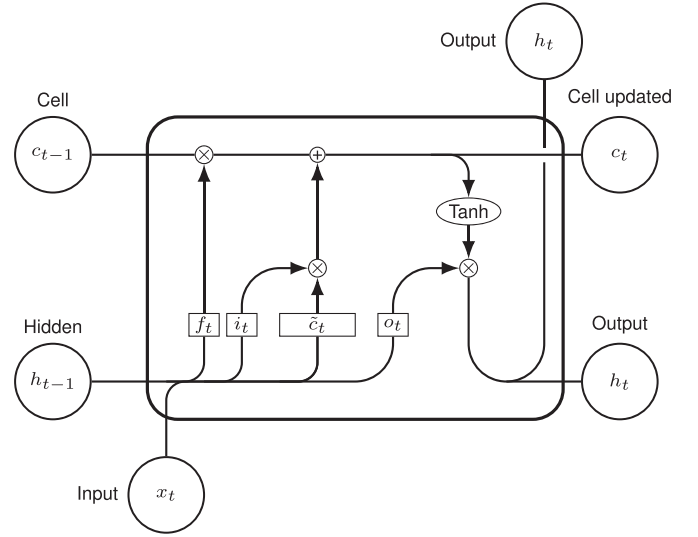


Fig. 2. Architecture of an LSTM cell. This takes into account the current word input, previous hidden state, and previous cell state to generate a new hidden and cell state. Connections between cells indicate the sequential flow of data through the model.

where C refers to the set of features obtained following the convolution operation. This approach retains the most relevant information while reducing dimensionality, thus facilitating subsequent processing in the network.

D. Block 2: BiLSTM Layer and Attention

The second block of our NLP model architecture comprises multiple layers, including a BiLSTM layer followed by a dropout layer, then another BiLSTM layer, and finally an attention layer. This configuration captures both short and long-term information in sequential data while focusing on the important parts of the sequence. The initial BiLSTM layer, featuring 64 units, captures contextual dependencies in both directions of the sequence. This bi-directionality allows the model to learn from contextual information both before and after each word, which is crucial for understanding the overall context of the text. A dropout layer with a rate of 0.2 is added to regularize the model and mitigate the risk of overfitting. This improves the model's generalization by making it less sensitive to noise in the training data.

The second BiLSTM layer, also with 64 units, continues learning contextual dependencies in the sequence, taking into account the representations learned by the first BiLSTM layer. This refines word representations based on the global context of the sequence. Finally, the attention layer is employed to assign weights to the importance of various parts of the sequence based on their relevance to the current task. This enables the model to focus on the most informative segments of the sequence when making decisions.

1) *BiLSTM Layer and Dropout*: The LSTM (Long Short-Term Memory) cell as shown in Fig. 2 is a basic unit used in LSTM networks, and the bidirectional version (BiLSTM) [24] consists of two LSTM cells: one processing the sequence in

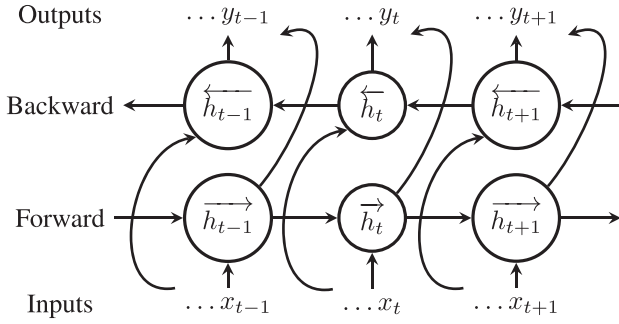


Fig. 3. BiLSTM architecture. This bidirectional architecture enables contextual information to be exploited in both temporal directions, simultaneously integrating past and future sequential representations. The arrows indicate the flow of information through the network layers, while the circles represent the hidden states of the BiLSTM model, with the upper part representing the hidden states before (Forward) and the lower part representing the hidden states after (Backward). Sequential inputs and outputs are also shown for each instant of time.

normal order and the other in reverse order. Letting x be the input at time t , h be the output at time t , c be the cell memory at time t , and $\odot$ be the Hadamard product (element by element), we obtain:

Forget gates (f_t):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (7)$$

Input gates (i_t):

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8)$$

Output gates (o_t):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

Cell candidate ($\tilde{c}_t$):

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (10)$$

Cell memory update (c_t):

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (11)$$

Update output (h_t):

$$h_t = o_t \odot \tanh(c_t) \quad (12)$$

In these equations, W_f, W_i, W_o, W_c represent the weights of the various gates in the LSTM cell, $[h_{t-1}, x_t]$ refer to the concatenation of the previous output vector and the current input, b_f, b_i, b_o, b_c are the corresponding biases, σ refers to the sigmoid function, and $\tanh$ to the hyperbolic tangent function.

The bidirectional LSTM (BiLSTM) [51] includes two sets of these equations - one for forward travel ($\vec{h}_t, \vec{c}_t$) and one for reverse travel ($\overleftarrow{h}_t, \overleftarrow{c}_t$) - as shown in Fig. 3. The final outputs H_t and C_t for the time step t are obtained by concatenating the outputs of the two directions:

$$H_t = [\vec{h}_t, \overleftarrow{h}_t] \quad (13)$$

$$C_t = [\vec{c}_t, \overleftarrow{c}_t] \quad (14)$$

This allows BiLSTM to take into account contextual information from both the past and future.

Dropout is a regularization technique that consists of randomly “dropping” (ignoring) certain neurons during training. This avoids overfitting by preventing the model from becoming too dependent on specific neurons. Mathematically, dropout is implemented by randomly assigning a probability p of zeroing to each neuron during each training run. The mathematical formula for the dropout operation on an input x with probability p is given by :

$$\text{dropout}(x) = \frac{\text{mask} \odot x}{1 - p} \quad (15)$$

where $\odot$ refers to the Hadamard product (element by element) and “mask” to a random binary mask with the same form as x , where probability p of each element is 0 and $(1-p)$ 1.

2) Attention Layer: Attention [25] is a technique that allows the model to focus on specific parts of the input sequence when making decisions. Within the context of NLP, this means that the model can place more importance on certain words or parts of the sequence during classification.

Attention is given by :

$$\text{Attention}(h_i) = \sum_j \alpha_{ij} \cdot h_j \quad (16)$$

where h_i refers to the BiLSTM output at position i in the sequence α_{ij} and to the attention weight associated with the pair (h_i, h_j) , and the sum is calculated over all positions j in the sequence.

Attention weights α_{ij} are generally calculated using the softmax function to ensure that the sum of the weights for a given position is equal to 1:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})} \quad (17)$$

where e_{ij} is a measure of the compatibility between positions i and j calculated, for example, by a scalar product between representations h_i and h_j .

The resulting weighted context c_i is then used to form the final contextual representation, which is in turn then concatenated with the BiLSTM output at position i to form the final representation of that position:

$$\text{Weighted context}(h_i) = \sum_j \alpha_{ij} \cdot h_j \quad (18)$$

$$\text{Final representation}(h_i) = [\text{Weighted context}(h_i); h_i] \quad (19)$$

Attention allows the model to place greater emphasis on certain parts of the sequence, depending on their importance to the task. Within the context of NLP, this can be particularly useful for processing text sequences of varying lengths by focusing on the most informative parts. Attention enhances the model’s ability to capture important relationships between words in a sequence and can thus lead to better performance in text classification tasks.

E. Concatenation of Parallel Layer Outputs

The integration of the outputs from both blocks is accomplished through a straightforward concatenation process, where

the output vectors from each block are joined side by side to form a unified representation. This method enables the amalgamation of insights derived from both blocks, allowing for a comprehensive understanding of the underlying text. Consequently, every component within the merged representation embodies a synthesis of the distinctive features extracted by the CNN block, which highlights localized and specific attributes of the text, and those extracted by the BiLSTM block with attention mechanism, adept at capturing prolonged dependencies and contextual nuances. This unified representation thus serves as a robust foundation for subsequent layers, facilitating a holistic analysis of the text's intricacies and facilitating enhanced predictive capabilities.

When we have outputs $X_{\text{CNN_maxpool}}$ of the CNN layer and $X_{\text{Attention}}$ of the BiLSTM layer with Attention, concatenation is simply an operation that joins these two vectors along a new dimension. The new concatenated representation, denoted as X_{concat} , is used as the input for dense layers, and this concatenation can be expressed as follows:

$$X_{\text{concat}} = \text{Concatenate}([X_{\text{CNN_maxpool}}, X_{\text{Attention}}]) \quad (20)$$

Where Concatenate refers to a function that takes vectors $X_{\text{CNN_maxpool}}$ and $X_{\text{Attention}}$ and joins them along a new dimension. Concatenation enables the model to benefit from information extracted by different architectures in parallel, thus improving its performance in terms of the biomedical text classification task.

F. Dense Layers

The dense layers in our architecture perform a non-linear transformation of the extracted features, resulting in a final output used for binary prediction in biomedical text classification. The first dense layer is a fully connected layer with 64 neural units and a ReLU activation function [59]. This dense layer captures complex relationships between different input data features, helping to create more abstract and meaningful representations. The ReLU function is used to introduce non-linearity into the model, effectively modeling complex phenomena and non-linear relationships between features. The second layer, also equipped with the ReLU function but with only 32 neural units, continues the feature distillation process by reducing the dimensionality of the feature space. This layer acts as a dimensionality reducer, emphasizing the most relevant and informative aspects of the data transformed by the first layer.

Finally, the third layer is a single-output layer with a single neural unit and a sigmoid activation function. This layer is responsible for binary classification of the data, producing a probabilistic output indicating the likelihood that the data belongs to one of the two classes. The sigmoid function is used because it effectively maps input values to a range between 0 and 1, allowing for a simple interpretation of the class probability.

III. EXPERIMENTAL SETUP

A. Dataset and Preprocessing

1) *Data Description*: In this study, we exploited the MIMIC IV [27] dataset to develop and validate our predictive model,

Yelp review Polarity dataset was used to validate the generalizability of the model. MIMIC IV, from the Beth Israel Deaconess Medical Centre, offers extensive temporal coverage from 2008 to 2019 and includes detailed data from over 60,000 patients.

2) *Data Retrieval*: Our initial step involved retrieving relevant medical reports from the MIMIC-IV databases. We employed Structured Query Language (SQL) joins within PostgreSQL to extract data from the “notes” table. Each document within this table is linked to a diagnostic code using the ICD-9 coding system (International Classification of Diseases version 9 Diagnostic). To isolate documents about lung cancer diagnoses and others, we implemented further SQL joins to match documents with lung cancer and other disease-related ICD-9 codes. The comprehensive summary in Table V in the appendix provides detailed information on the diseases considered in this dataset, along with the number of documents associated with each disease.

3) *Data Reconstruction*: Given that each document encompasses multiple paragraphs delineating diverse facets of patient medical history and treatment, encompassing admission details (medical history, social history, current condition, administered medications, clinical examinations, and radiology reports), we conducted data filtering to isolate pertinent information. Our focus was specifically directed towards extracting details pertinent to medical history, social history, and familial history. We also took into account the patient's present illness to establish connections with historical information. All this information was selected in each patient's document. Patient medical history plays a pivotal role in assessing potential cancer risk through an understanding of past medications and documented disorders. Moreover, social history information enables the discernment of smoking and alcohol consumption patterns, facilitating the correlation between these lifestyle factors and the data gleaned from medical and family histories, thereby aiding in the prediction of lung cancer risk.

4) *Data Cleaning*: Subsequently, text cleaning was performed to remove special characters and dates from the datasets, resulting in the final datasets for experimentation.

5) *Data Selection*: The use of MIMIC IV to train our model was motivated by its qualitative improvements. However, to guarantee the robustness and generalisability of our model, we opted to make an experiment using the second dataset Yelp review polarity. This approach enabled us to assess the performance of our model on data from a distinct context, to confirm the robustness of the model to be generalizable in other databases. The MIMIC IV database was divided into three distinct sets for training, validation, and test sets, representing 70%, 15%, and 15% of total data, respectively. Specifically, the training set comprises 26,807 documents, while the validation and test set each comprises 5,745 documents, thus totaling 38,297 documents in the case of the MIMIC IV set. The MIMIC IV dataset provides two distinct classes: class 0, corresponding to non-cancerous individuals, and class 1, grouping patients diagnosed with lung cancer. This stratification resulted in a balanced dataset, comprising 19,150 instances in the case of class 0 and 19,147 instances in that of class 1. These details are crucial to interpreting the results, ensuring a thorough analysis

of the model's ability to effectively distinguish between the two classes within a variety of medical data contexts.

B. Hyperparameter Setting

The hyperparameters selected in this experiment are described in detail below. Words in vector space are embedded with a dimension of 100, and the Conv1D layer uses 128 filters with a convolution window of size 5 to extract local features. For its part, the Bidirectional LSTM layer, size 64, enables the bidirectional modeling of sequences, and a dropout rate of 0.2 is then applied after this layer to prevent overfitting. The dense layers that follow combine features extracted by CNN and Bi-LSTM, with two layers of 64 units each, activated by ReLU. Finally, the output layer, with one unit and a sigmoid activation function, performs the binary classification. The Adam optimizer is used with 0.9 in the case of beta_1 (the exponential flattening coefficient for the first-moment estimate in Adam.) and 0.999 in that of beta_2 (the exponential flattening coefficient for the second-moment estimate in Adam.) - with 10 epochs, a batch size of 32, and learning rate of 0.001.

C. Evaluation Metrics

As part of the evaluation of our model, several metrics [48], [49], [50] were used to assess its performance in binary classification. These metrics include accuracy, recall, precision, F1-score, and AUC-ROC, [28] each providing a distinct perspective on the quality of the model's ability for prediction.

- Accuracy: measures the overall rate of correct predictions by the model. Mathematically, it is defined as :

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (21)$$

- Recall (Sensitivity): measures the model's ability to identify all positive examples, calculated by the formula :

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (22)$$

- Precision: measures the accuracy of positive predictions made by the model, defined by the equation :

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (23)$$

- F1-score: The F1-score is a metric that combines recall and precision into a single value, calculated by the formula :

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (24)$$

- AUC-ROC (Area Under the Receiver Operating Characteristic curve): AUC-ROC measures the model's ability to discriminate between positive and negative classes at different classification thresholds, with a higher AUC-ROC indicating better discrimination ability. AUC-ROC is calculated by integrating the ROC curve, which plots the true-positive rate against the false-positive rate.
- The Matthews Correlation Coefficient (MCC): is a robust metric used to evaluate the performance of binary classification models. Unlike accuracy, which can be misleading in imbalanced datasets, MCC takes into account

true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for a more comprehensive assessment.

$$\text{MCC} = \frac{(\text{TP} \cdot \text{TN}) - (\text{FP} \cdot \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}}$$

In addition to the metrics described above, we adopted another rigorous approach to evaluate the performance of our model, using the Stratified K-Fold Cross-Validation strategy. [29] This method played a crucial role in the robustness of the evaluation by dividing the dataset into K folds [46], [47] while keeping class distribution in each fold similar to that of the original dataset. More specifically, we chose K equal to 5, meaning that the dataset was divided into five distinct folds. During each iteration of the cross-validation process, the model was trained on four folds and evaluated on the remaining fold. This procedure was repeated five times, so that each fold was used once as a validation set, ensuring that each sample was included exactly once in the validation set. The use of stratified cross-validation is particularly crucial within the context of our binary classification task, where it is essential to ensure that each fold retains the same proportion of positive and negative classes as the original data set. This mitigates the risk of bias resulting from imbalances in class distribution. By adopting this approach, our model evaluation becomes more robust, as it enables performance to be more reliably generalized over the full dataset. This cross-validation method reinforces the credibility of the results obtained and ensures a fair assessment of the model's ability to generalize new data, thereby strengthening the validity and reliability of our conclusions.

IV. RESULTS

This section presents the key findings from a comprehensive evaluation of our proposed binary classification model. The evaluation compared our model's performance to established benchmarks, including BioBERT and LSTM models [52]. We employed the MIMIC IV test set for the primary evaluation. Additionally, to assess the model's generalizability, we conducted a secondary evaluation using the Yelp review polarity dataset, comparing our model to state-of-the-art models trained on this dataset. These results shed light on the robustness and overall effectiveness of our approach in the context of lung cancer detection.

A. Evaluation Performance on MIMIC IV Test Set

Table I shows the performance results of the models evaluated on the test set of the MIMIC IV dataset, with a particular focus on lung cancer detection. Our proposed model obtained outstanding results, with precision, recall, and F1-score all equal to 98.0%, thus demonstrating its effectiveness in classifying positive and negative cases. The accuracy and AUC-ROC of our model also attained a remarkable 98.1% and 100.0% respectively, confirming its ability to discriminate effectively between classes.

In comparison, the LSTM model performed strongly, albeit slightly worse than our proposal, with an F1-score of 97%.

TABLE I
PERFORMANCE ON MIMIC IV TEST SET

Model	Confusion matrix			Performance evaluation metrics (%)						Dataset size	
	Predictive class →			Precision	Recall	F1-score	Accuracy	AUC	MCC		Test size
	Actual class ↓	0	1								
Proposed model	0	2836	51	98.0	98.0	98.0	98.1	100.0	96.2		2887
	1	58	2800	98.0	98.0	98.0					
LSTM	0	2852	35	95.0	99.0	97.0	97.0	99.0	93.5		2887
	1	153	2705	99.0	95.0	97.0					
BioBERT	0	2819	68	98.0	98.0	98.0	98.0	98.0	95.5		2887
	1	61	2797	98.0	98.0	98.0					

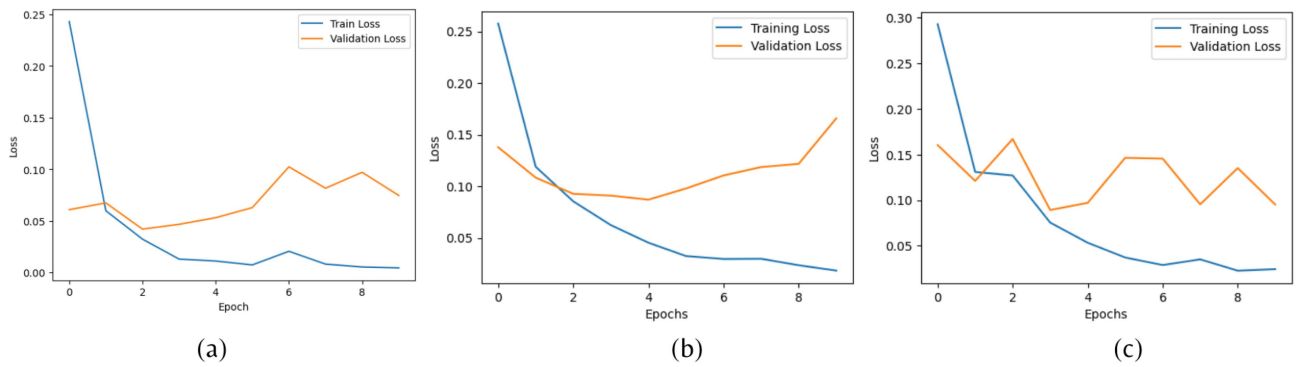


Fig. 4. This figure presents the learning curves of our experiment using three different models: our proposed CNN-BiLSTM Attention model (a), the LSTM model (b), and the BioBERT model (c). Each subplot visualizes the error rates during the training phase, highlighting the distinct learning trajectories of each model.

BioBERT, although performing well, also obtained similar results to our model in terms of precision F1-score and Recall, underlining the competitiveness of our approach. In terms of MCC, our model performs the two comparative models, with 96.2% of MCC. We assessed the effectiveness of our suggested model in predicting class 1, which corresponds to persons with cancer, by comparing it to the BioBERT and LSTM models. The findings, as shown in Table I, provide valuable insights into the resilience of our model. Upon studying the confusion matrix, it became evident that our model achieved a significant number of true positives (TP), indicating its proficiency in accurately identifying instances of cancer. Fig. 4 shows the training curves in terms of errors of the models compared, notably our model, the LSTM, and BioBERT.

B. Evaluation Performance on MIMIC IV Using Stratified Cross-Validation

To guarantee the robustness and generalizability of our proposed CNN-BiLSTM attention model for lung cancer detection, we employed stratified 5-fold cross-validation. This rigorous evaluation methodology involves partitioning the MIMIC IV dataset into five stratified folds, ensuring that the proportion of positive and negative lung cancer cases remains consistent within each fold. This approach effectively mitigates the risk of overfitting, a common challenge in machine learning, by

preventing the model from memorizing specific patterns in the training data that may not generalize well to unseen data.

Our proposed model consistently outperformed established benchmarks, including Long Short-Term Memory (LSTM) and BioBERT, across a comprehensive range of performance metrics. Notably, it achieved an outstanding average accuracy, recall, and F1-score of 98.4%. This remarkable consistency across different data subsets within the MIMIC IV dataset underscores the model's ability to generalize effectively and accurately identify lung cancer cases from medical notes.

The model's performance is further corroborated by its impressive area under the receiver operating characteristic curve (AUC-ROC) of 99.8%. This metric indicates the model's superior ability to distinguish between relevant medical text classes, particularly positive and negative lung cancer cases. This success can be attributed to the synergistic combination of Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory Networks (BiLSTMs) within our model. Detailed results are presented in Table II.

C. Evaluation Performance on Yelp Review Polarity Dataset

We conducted an additional evaluation using the Yelp Review Polarity dataset. Here, our model again displayed outstanding performance, achieving an average Precision, Recall, and

TABLE II
AVERAGE PERFORMANCE ON MIMIC IV DATASET USING STRATIFIED K CROSS VALIDATION

Model	Class label	Precision $\pm$ std	Recall $\pm$ std	F1-score $\pm$ std	Accuracy $\pm$ std	AUC-ROC $\pm$ std
Proposed model	0	98.2 $\pm$ 0.0040	98.6 $\pm$ 0.0048	98.4 $\pm$ 0.0048	98.4 $\pm$ 0.0048	99.8 $\pm$ 0.0001
	1	98.6 $\pm$ 0.0048	98.2 $\pm$ 0.0040	98.4 $\pm$ 0.0048		
	Average	98.4 $\pm$ 0.0044	98.4 $\pm$ 0.0044	98.4 $\pm$ 0.0048		
LSTM	0	97.8 $\pm$ 0.0040	89.6 $\pm$ 0.0233	93.2 $\pm$ 0.0132	93.5 $\pm$ 0.0101	97.9 $\pm$ 0.0030
	1	90.4 $\pm$ 0.0224	97.8 $\pm$ 0.0040	93.8 $\pm$ 0.0116		
	Average	94.1 $\pm$ 0.0132	93.7 $\pm$ 0.0136	93.4 $\pm$ 0.0134		
BioBERT	0	97.6 $\pm$ 0.0185	98.8 $\pm$ 0.0040	98.4 $\pm$ 0.0040	98.2 $\pm$ 0.0116	98.3 $\pm$ 0.0095
	1	98.8 $\pm$ 0.0040	97.6 $\pm$ 0.0185	98.2 $\pm$ 0.0116		
	Average	98.2 $\pm$ 0.0112	98.2 $\pm$ 0.0112	98.3 $\pm$ 0.0098		

TABLE III
PERFORMANCE ON YELP REVIEW POLARITY DATASET

Confusion matrix			Performance evaluation metrics (%)						Dataset size
Predictive class $\rightarrow$			Precision	Recall	F1-score	Accuracy	AUC	Matthews (MCC)	Test size
Actual class $\downarrow$	0	1							
0	18141	859	94.9	95.4	95.2	95.1	99.0	90.3	19000
1	968	18032	95.4	94.9	95.1				19000
Average			95.1	95.1	95.1				

F1-score of 95.1% each, alongside an Accuracy of 95.1% and an Area Under the ROC Curve (AUC) of 99.0%. Notably, the MCC reached 90.3%, highlighting the model's strong ability to distinguish between positive and negative sentiment classes in a non-medical context. These findings strongly suggest that the proposed CNN-BiLSTM attention model offers exceptional promise for lung cancer detection in real-world scenarios. The model's generalizability, showcased through its performance on a broader text classification task, further strengthens its potential for wider application. Table III summarize the result of our model on Yelp review polarity dataset.

V. DISCUSSION

A. Generalizability

To further assess the generalizability of our proposed CNN-BiLSTM attention model, we conducted an evaluation using the Yelp Review Polarity dataset, a widely recognized benchmark for text classification tasks. This dataset consists of customer reviews labelled as positive or negative. The performance of our model surpasses that of most robust transformer-based models in this dataset. However, it is essential to acknowledge that a few models achieve slightly better results [62]. A significant advantage of our model lies in its efficient architecture, featuring only 12.5 million trainable parameters compared to model like KEN-BLOOM (531 million) [62]. This compact size plays a crucial role in facilitating practical deployment scenarios and enhances the model's interpretability. The reduced number of parameters [66] not only makes the model more feasible for

TABLE IV
COMPARISON WITH PREVIOUS RESEARCH ON YELP REVIEW POLARITY DATASET (%)

Model	Year	Trainable Parameters	Acc
CNN-Fast text [61]	2023	-	95.0
Distillation [63]	2021	-	92.1
KEN - BERT [62]	2024	57 M	93.8
KEN - DistilBERT [62]	2024	40 M	94.1
KEN - Electra [62]	2024	14 M	93.7
KEN - BLOOM [62]	2024	531 M	96.3
LSTM [62]	2023	-	91.0
Proposed model		12.5 M	95.1

real-world applications with limited computational resources but also allows for greater transparency and understanding of the model's decision-making processes. This balance of high performance and efficiency underscores the potential of our model in diverse practical contexts. [66] The results achieved by state-of-the-art models in comparison with our model are listed in Table IV.

This observation highlights the model's ability to achieve excellent performance on diverse tasks while maintaining a degree of parameter efficiency. This characteristic is crucial

for real-world applications, where computational resources and deployment considerations are often paramount.

The successful transferability of our model from the medical domain (MIMIC IV) to the general text classification domain (Yelp reviews) underscores its potential for broader applicability. While certain models in Table III achieve slightly higher accuracy on the Yelp review dataset, it's important to consider the trade-off between performance and model complexity. Our model demonstrates strong generalizability with a smaller footprint, making it a compelling choice for various text classification tasks where both accuracy and efficiency are desired.

B. Explainability

To gain deeper insights into the inner workings of our CNN-BiLSTM attention model and enhance its interpretability, we employed the SHAP (SHapley Additive exPlanations) values approach [65]. SHAP values provide a unique way to explain the predictions of complex machine learning models by attributing the contribution of each input feature to the overall prediction.

In the context of our lung cancer detection model, SHAP values allow us to understand how specific words or phrases in medical notes influence the model's decision to classify a patient as having or not having lung cancer. By examining the SHAP values for each input [67], we can identify the most important features that contribute to the model's predictions.

Our analysis of SHAP values revealed a fascinating pattern of terms that consistently hold high positive SHAP values for predicting lung cancer cases. These terms include:

“smoker,” “cancer,” “carcinoma,” “metastatic,” “nonsmall,” “cell,” “recently,” “ca” These terms are not surprising, as they are directly related to the diagnosis and characteristics of lung cancer. Their presence in medical notes strongly suggests the possibility of lung cancer, and our model effectively learns to identify these patterns to make accurate predictions.

On the other hand, terms like “vomiting,” “day,” “mi,” “old,” “lung,” “noted,” “etoh,” “mr,” “ascites,” “found,” “lobe” often exhibit important SHAP values. While these terms may still be relevant to a patient's medical history, they are less directly indicative of lung cancer and therefore contribute less to the model's predictions.

The interpretability provided by SHAP values not only enhances our understanding of the model's decision-making process but also offers valuable insights for improving its performance. [68] By identifying the most important features, we can focus our data collection and preprocessing efforts on ensuring the availability of high-quality data for these critical elements.

Moreover, SHAP values can be used to identify potential biases or blind spots in the model. For instance, if certain terms consistently exhibit negative SHAP values, it may indicate that the model is underweighting those features, potentially leading to misclassifications. This Fig. 5 illustrates the SHAP value of 100 randomly selected instances.

In the realm of machine learning, understanding the factors that drive a model's predictions is crucial for building trust and ensuring its reliability [69]. In the context of our lung cancer detection model, deciphering the role of specific words

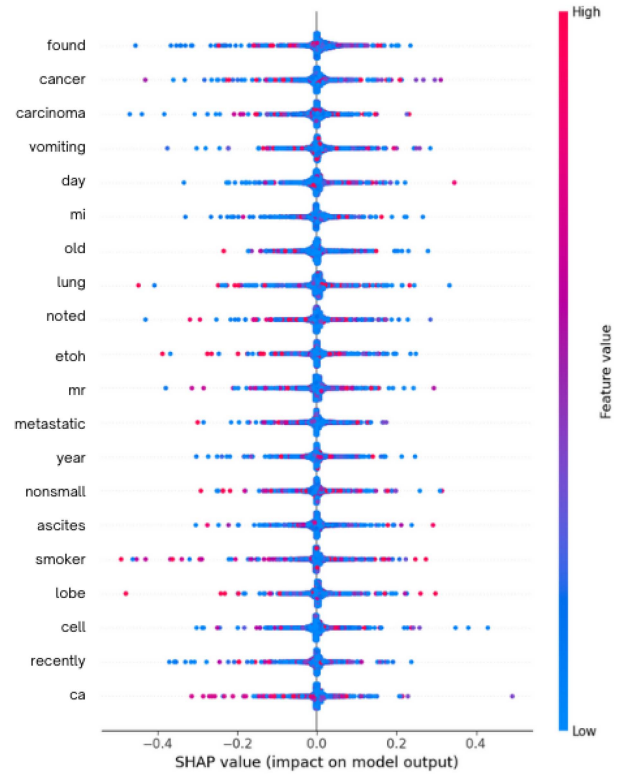


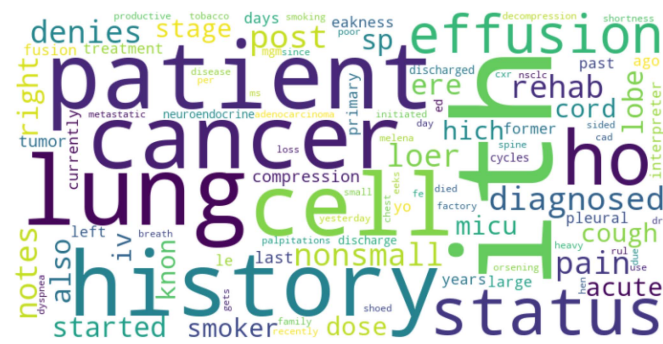
Fig. 5. This figure illustrates the SHAP value analysis for 100 randomly selected instances, highlighting the importance of individual words in the model's predictions. Each word in the figure contributes to the model's decision-making process, with the color and position indicating its significance. Visually, it is evident that most words share similar importance levels, although minor variations are noticeable. This uniformity in word importance reinforces the model's robustness and reliability, enhancing confidence in its ability to generalize and accurately predict new examples.

and phrases in medical notes is essential for comprehending its decision-making process and identifying potential areas for improvement. While word clouds and SHAP values offer distinct approaches to analyzing feature importance, their complementary strengths can be harnessed to gain a deeper understanding of our model's inner workings.

Word clouds provide a visually [70] intuitive representation of feature importance, where larger words indicate a higher perceived significance. In Fig. 6, we observe terms like “cancer”, “lung” and “cell” dominating the word cloud, suggesting their prominent role in the model's predictions. Others words like “smoker,” “lobe” and “nonsmall” have also a mean importance in the word cloud visualization. This aligns with our medical expertise, as these terms are directly associated with lung cancer diagnosis and characteristics.

The combination of word clouds and SHAP values offers a comprehensive approach to understanding feature importance. [71] Word clouds provide a qualitative overview of high-level importance patterns, while SHAP values deliver quantitative measures of feature contributions for individual predictions.

The proposed CNN-BiLSTM Attention model stands out due to its unique architectural features [72] and the significant



advantages it offers. The model effectively extracts features through the combination of CNN and BiLSTM layers, capturing both local and contextual features from clinical notes, which allows it to identify fine-grained patterns and long-range dependencies within the text. The attention mechanism dynamically focuses on the most relevant parts of the input sequence, enabling the model to assign higher weights to informative phrases and sentences that are crucial for accurate classification. This attention mechanism also adds a degree of interpretability to the model, as the attention weights provide insights into the model's decision-making process, revealing which parts of the clinical notes contribute most to the classification outcome.

We strictly adhered to ethical guidelines to ensure the protection of individuals' privacy, safety, and well-being throughout data collection, processing, and analysis. The MIMIC IV database complies with ethical standards and anonymous personal information to protect patient privacy. Relevant clinical notes on lung cancer diagnosis were carefully selected to minimize exposure to sensitive data.

institutions, AI developers, and regulatory bodies is essential to address these concerns and facilitate adoption. Practical steps for integration include conducting pilot studies, refining the model based on clinician feedback, and establishing protocols for data collection and interpretation. all those recommendations need to be considered before any integration into clinical workflows.

VI. CONCLUSION

The proposed model leverages the power of deep learning to effectively extract meaningful patterns from medical notes in the MIMIC IV dataset. Our comprehensive evaluation demonstrates the model’s superior performance compared to state-of-the-art approaches like BioBERT and LSTM. Notably, our model achieves a competitive accuracy of 98.1% with a significantly smaller number of trainable parameters (12.5 million) compared to BioBERT’s 110 million parameters. This highlights the model’s efficiency and parameter optimization capabilities.

To enhance interpretability and gain deeper insights into the model's decision-making process, we employed the SHAP (Shapley Additive exPlanations) values approach. SHAP values reveal the contribution of each word in a medical note to the model's prediction, providing valuable insights into the factors that influence the model's classification.

Our CNN-BiLSTM attention model stands as a promising contribution to the fight against lung cancer. Its superior performance, efficiency, and generalizability make it a compelling tool for clinical decision support. As we continue to explore its potential and expand its scope, we are confident that this model can play a significant role in improving patient outcomes and saving lives.

TABLE V
DISEASES SELECTION DETAILS

ICD-9	Description	MIMIC IV
Lung cancer		Number of documents
162.0	Malignant neoplasm of trachea	312
162.2	Malignant neoplasm of main bronchus	621
162.3	Malignant neoplasm of upper lobe, bronchus or lung	4729
162.4	Malignant neoplasm of middle lobe, bronchus or lung	409
162.5	Malignant neoplasm of lower lobe, bronchus or lung	2476
162.8	Malignant neoplasm of other parts of bronchus or lung	2970
162.9	Malignant neoplasm of bronchus and lung, unspecified	7630
Sub total		19147
Other diseases		
155.0	Malignant neoplasm of liver, primary	10493
155.1	Malignant neoplasm of intrahepatic bile ducts	2387
170.7	Malignant neoplasm of long bones of lower limb	1434
153.9	Malignant neoplasm of colon, unspecified site	1183
153.3	Malignant neoplasm of sigmoid colon	803
153.6	Malignant neoplasm of ascending colon	541
153.4	Malignant neoplasm of cecum	488
155.2	Malignant neoplasm of liver, not specified as primary or secondary	457
153.8	Malignant neoplasm of other specified sites of large intestine	291
153.2	Malignant neoplasm of descending colon	285
153.1	Malignant neoplasm of transverse colon	195
153.0	Malignant neoplasm hepatic flexure	115
170.6	Malignant neoplasm of pelvic bones, scrum and coccyx	93
153.7	Malignant neoplasm of splenic flexure	91
153.5	Malignant neoplasm of appendix vermiformis	76
175.9	Malignant Malignant neoplasm of other and unspecified sites of male breast	73
170.9	Malignant neoplasm of bone and articular cartilage, site unspecified	36
170.4	Malignant neoplasm of scapula and long bones of upper limb	26
170.2	Malignant neoplasm of vertebral column, excluding sacrum and coccyx	20
170.0	Malignant neoplasm of bones of skull and face, except mandible	19
170.3	Malignant neoplasm of ribs, sternum, and clavicle	18
175.0	Malignant neoplasm of nipple and areola of male breast	16
170.1	Malignant neoplasm of mandible	6
170.5	Malignant neoplasm of short bones of upper limb	4
Total		38297

APPENDIX

In the scope of our study, we conducted a meticulous selection of textual data for classification, focusing on medical documents related to lung cancer as well as other types of diseases. In this appendix, we provide detailed information on the diseases considered in both datasets. For lung cancer classification, we identified several diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD-9) [57] corresponding to different locations and types of lung cancer. The selected diseases include malignant neoplasms affecting various parts of the respiratory tract, including the trachea, main bronchi, upper, middle, and lower lobes of the bronchi or lungs, as well as other parts of the bronchi and lungs. These diseases, represented by their respective ICD-9 codes, were carefully included in our analysis to ensure a comprehensive representation of lung cancer cases in the datasets. Simultaneously, other diseases were also considered to form a distinct class in our dataset. These diseases encompass various types of malignant neoplasms affecting different parts of the body, such as the liver, colon, rectum, breast, bones, and other anatomical sites. Each disease was associated with its specific ICD-9 code, and the corresponding number of

documents in the MIMIC IV datasets was recorded. In total, our dataset comprises a significant number of documents, with a total of 38,297 documents covering a wide range of diseases relevant to our study.

ACKNOWLEDGMENT

We wish to thank the eVida Research Group at the University of Deusto, recognized by the Basque Government under code IT1536-22, for its untiring support and commitment to providing us with the resources necessary to carry out the study until the time of its completion.

REFERENCES

- [1] A. H. Krist et al., "Screening for lung cancer: US preventive services task force recommendation statement," *Jama*, vol. 325, no. 10, pp. 962–970, 2021.
- [2] A. R. W. Sait, "Lung cancer detection model using deep learning technique," *Appl. Sci.*, vol. 13, no. 22, 2023, Art. no. 12510, doi: [10.3390/app132212510](https://doi.org/10.3390/app132212510).
- [3] K. L. Kehl et al., "Natural language processing to ascertain cancer outcomes from medical oncologist notes," *JCO Clin. Cancer Inform.*, vol. 4, pp. 680–690, Aug. 2020, doi: [10.1200/CCI.20.00020](https://doi.org/10.1200/CCI.20.00020).

- [4] Z. Gandhi et al., "Artificial intelligence and lung cancer: Impact on improving patient outcomes," *Cancers*, vol. 15, no. 21, Oct. 2023, Art. no. 5236, doi: [10.3390/cancers15215236](https://doi.org/10.3390/cancers15215236).
- [5] N. Sverzellati et al., "Low-dose computed tomography for lung cancer screening: Comparison of performance between annual and biennial screen," *Eur. Radiol.*, vol. 26, pp. 3821–3829, 2016.
- [6] H. L. Lancaster, M. A. Heuvelmans, and M. et Oudkerk, "Low-dose computed tomography lung cancer screening: Clinical evidence and implementation research," *J. Intern. Med.*, vol. 292, no. 1, pp. 68–80, 2022.
- [7] P. V. Kukhareva et al., "Inaccuracies in electronic health records smoking data and a potential approach to address resulting underestimation in determining lung cancer screening eligibility," *J. Amer. Med. Informat. Assoc.*, vol. 29, no. 5, pp. 779–788, 2022.
- [8] L. A. Cannon-Albright, S. R. Carr, and W. Akerley, "Population-based relative risks for lung cancer based on complete family history of lung cancer," *J. Thoracic Oncol.*, vol. 14, no. 7, pp. 1184–1191, 2019.
- [9] K. Kowsari et al., "Text classification algorithms: A survey," *Information*, vol. 10, 2019, Art. no. 150.
- [10] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer classification using Gaussian naive bayes algorithm," in *2019 Int. Eng. Conf. (IEC)*, 2019, pp. 165–170.
- [11] K. J. Dsouza and Z. Ansari, "Experimental exploration of support vector machine for cancer cell classification," in *2017 IEEE Int. Conf. Cloud Comput. Emerg. Markets*, 2017, pp. 29–34.
- [12] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021.
- [13] M. Botlagunta et al., "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Sci. Rep.*, vol. 13, no. 1, 2023, Art. no. 485.
- [14] S. Mithun et al., "Clinical concept-based radiology reports classification pipeline for lung carcinoma," *J. Digit. Imag.*, vol. 36, pp. 812–826, 2023.
- [15] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.
- [16] J. Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [17] E. Yang et al., "Transformer versus traditional natural language processing: How much data is enough for automated radiology report classification?," *Brit. J. Radiol.*, vol. 96, 2023, Art. no. 20220769.
- [18] D. Edara et al., "Sentiment analysis and text categorization of cancer medical records with LSTM," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 5, pp. 5309–5325, 2023.
- [19] P. M. Lavanya and E. Sasikala, "IFM-RCNN: A hybrid text classifier with enhanced performance of binary drug classification from tweets using improved faster mask-current convolutional neural network," *Knowl. Inf. Syst.*, vol. 66, pp. 557–579, 2023.
- [20] S. Zaman et al., "Automatic diagnosis labeling of cardiovascular MRI by using semisupervised natural language processing of text reports," *Radiol.: Artif. Intell.*, vol. 4, no. 1, 2021, Art. no. e210085.
- [21] T. Mikolov et al., "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
- [22] T. Mikolov et al., "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, vol. 26.
- [23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [24] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
- [25] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.
- [26] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, 2016, Art. no. 160035.
- [27] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Sci. Data*, vol. 10, no. 1, 2023, Art. no. 1.
- [28] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Mining Knowl. Manage. Process.*, vol. 5, no. 2, p. 1, 2015.
- [29] B. Efron and R. Tibshirani, "Improvements on cross-validation: The 632 bootstrap method," *J. Amer. Stat. Assoc.*, vol. 92, no. 438, pp. 548–560, 1997.
- [30] NIU et al., "Enhancing local feature extraction with global representation for neural text classification," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 496–506.
- [31] J. Deng, L. Cheng, and Z. Wang, "Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification," *Comput. Speech Lang.*, vol. 68, 2021, Art. no. 101182.
- [32] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [33] D. Zhang et al., "Combining structured and unstructured data for predictive models: A deep learning approach," *BMC Med. Informat. Decis. Mak.*, vol. 20, no. 1, pp. 1–11, 2020.
- [34] S. R. Stahlschmidt et al., "Multimodal deep learning for biomedical data fusion: A review," *Brief. Bioinf.*, vol. 23, no. 2, 2022, Art. no. bbab569.
- [35] L. R. Soenksen et al., "Integrated multimodal artificial intelligence framework for healthcare applications," *NPJ Digit. Med.*, vol. 5, no. 1, 2022, Art. no. 149.
- [36] J. Liu et al., "Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities," *J. Biomed. Informat.*, vol. 145, 2023, Art. no. 104466.
- [37] A. V. Singh et al., "Integrative toxicogenomics: Advancing precision medicine and toxicology through artificial intelligence and OMICs technology," *Biomed. Pharmacother.*, vol. 163, 2023, Art. no. 114784.
- [38] A. Kline et al., "Multimodal machine learning in precision health: A scoping review," *npj Digit. Med.*, vol. 5, no. 1, 2022, Art. no. 171.
- [39] C. Piao et al., "Word-context attention for text representation," *Neural Process. Lett.*, vol. 55, no. 8, pp. 11721–11738, 2023.
- [40] Q. Umer, H. Liu, and I. Illahi, "CNN-based automatic prioritization of bug reports," *IEEE Trans. Rel.*, vol. 69, no. 4, pp. 1341–1354, Dec. 2020.
- [41] A. J. Thirunavukarasu et al., "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.
- [42] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.
- [43] N. Ali et al., "Enhancing the performance of cancer text classification model based on cancer hallmarks," *IAES Int. J. Artif. Intell.*, vol. 10, no. 2, pp. 316–323, 2021.
- [44] S. Gao et al., "Limitations of transformers on clinical text classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3596–3607, Sep. 2021.
- [45] F. Gargiulo et al., "Deep convolution neural network for extreme multi-label text classification," in *Proc. Healthinf*, 2018, pp. 641–650.
- [46] T. R. Mahesh et al., "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, 2023, Art. no. 100247.
- [47] S. Prusty et al., "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Front. Nanotechnol.*, vol. 4, 2022, Art. no. 972421.
- [48] H. Ogura et al., "Comparison of metrics for feature selection in imbalanced text classification," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 4978–4989, 2011.
- [49] S. Minaee et al., "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surv. (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.
- [50] M. Ikonmakis et al., "Text classification using machine learning techniques," *WSEAS Trans. Comput.*, vol. 4, no. 8, pp. 966–974, 2005.
- [51] J. Zhang et al., "LSTM-CNN hybrid model for text classification," in *2018 IEEE 3rd Adv. Inf. Technol. Electron. Automat. Control Conf.*, 2018, pp. 1675–680.
- [52] X. Bai, "Text classification based on LSTM and attention," in *Proc. IEEE 13th Int. Conf. Digit. Inf. Manage. (ICDIM)*, 2018, pp. 29–32.
- [53] W. Yin et al., "Comparative study of cnn and rnn for natural language processing," 2017, *arXiv:1702.01923*.
- [54] S. Soni, S. S. Chouhan, and S. S. Rathore, "TextConvoNet: A convolutional neural network based architecture for text classification," *Appl. Intell.*, vol. 53, pp. 14249–14268, 2023, doi: [10.1007/s10489-022-04221-9](https://doi.org/10.1007/s10489-022-04221-9).
- [55] L. Xiaoyan, R. C. Raga, and S. Xuemei, "GloVe-CNN-BiLSTM model for sentiment analysis on text reviews," *J. Sensors*, vol. 2022, 2022, Art. no. 7212366, doi: [10.1155/2022/7212366](https://doi.org/10.1155/2022/7212366).
- [56] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, "LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, 2020, Art. no. 5841, doi: [10.3390/app10175841](https://doi.org/10.3390/app10175841).
- [57] C. for Medicare and M. Services, "ICD-9-CM diagnosis and procedure codes: Abbreviated and full code title," [Online]. Available: <https://www.cms.gov/medicare/codingbilling/icdclcodes/icd-cmiagnosisprocedurecodesabbreviatedandfullcodetitles>. visited 05/04/2024
- [58] N. Norori, Q. Hu, F. M. Aellen, D. F. Faraci, and A. Tzovara, "addressing bias in big data and AI for health care: A call for open science," *Patterns*, vol. 2, 2021, Art. no. 100347, doi: [10.1016/j.patter.2021.100347](https://doi.org/10.1016/j.patter.2021.100347).
- [59] A. Agarap, "Deep learning using rectified linear units (relu)," 2018, *arXiv:1803.08375*.

- [60] F. Gama, D. Tyskbo, J. Nygren, J. Barlow, and J. Reed, "Svedberg p implementation frameworks for artificial intelligence translation into health care practice: Scoping," *Rev. J. Med. Internet Res.* vol. 24, no. 1, 2022, Art. no. e32215, doi: [10.2196/32215](https://doi.org/10.2196/32215).
- [61] M. Umer et al., "Impact of convolutional neural network and FastText embedding on text classification," *Multimedia Tools Appl.*, vol. 82, no. 4, pp. 5569–5585, 2023.
- [62] M. Mastromattei and F. M. Zanzotto, "Less is KEN: A universal and simple non-parametric pruning algorithm for large language models," 2024, *arXiv:2402.03142*.
- [63] Y. Li and W. Li, "Data distillation for text classification," 2021, *arXiv:2104.08448*.
- [64] M. Sadikin and A. Fauzan, "Evaluation of machine learning approach for sentiment analysis using yelp dataset," *Eur. J. Elect. Eng. Comput. Sci.*, vol. 7, no. 6, pp. 58–64, 2023.
- [65] H. Kaur et al., "Interpreting interpretability: Understanding data scientist use of interpretability tools for machine learning," in *Proc. 2020 CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–14.
- [66] M. Denil et al., "Predicting parameters in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, vol. 26.
- [67] J. Mullenbach et al., "Explainable prediction of medical codes from clinical text," 2018, *arXiv:1802.05695*.
- [68] M. Nazar et al., "A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques," *IEEE Access*, vol. 9, pp. 153316–153348, 2021.
- [69] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.
- [70] J. Li et al., "Visualizing and understanding neural models in NLP," 2015, *arXiv:1506.01066*.
- [71] T. Adel, Z. Ghahramani, and A. Weller, "Discovering interpretable representations for both deep generative and discriminative models," in *Int. Conf. Mach. Learn.*, PMLR, 2018, pp. 50–59.
- [72] L. Yao, C. Mao, and Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," *BMC Med. Inform. Decis. Mak.*, vol. 19, 2019, pp. 31–39.
- [73] E. S. Alamoudi and N. S. Alghamdi, "Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings," *J. Decis. Syst.*, vol. 30, no. 2, pp. 259–281, 2021.