IEEE Journal of Translational Engineering in Health and Medicine

Received 2 October 2024; revised 5 January 2025; accepted 22 January 2025. Date of publication 28 January 2025; date of current version 13 February 2025.

Digital Object Identifier 10.1109/JTEHM.2025.3535676

Multi-Branch CNN-LSTM Fusion Network-Driven System With BERT Semantic Evaluator for Radiology Reporting in Emergency Head CTs

SELENE TOMASSINI[1], (Member, IEEE), DAMIANO DURANTI[1], ABDALLAH ZEGGADA[1], CARLO COSIMO QUATTROCCHI[2], FARID MELGANI[1], (Fellow, IEEE), AND PAOLO GIORGINI[1]

[1]Department of Information Engineering and Computer Science, University of Trento, 38121 Trento, Italy
[2]Centre for Medical Sciences, University of Trento, 38121 Trento, Italy

(C. C. Quattrocchi, F. Melgani, and P. Giorgini contributed equally to this work.)
CORRESPONDING AUTHOR: S. TOMASSINI (selene.tomassini@unitn.it)

This work was supported in part by the Piano Nazionale di Ripresa e Resilienza (PNRR) project FAIR-Future Artificial Intelligence Research and in part by the Ministero dell'Università e della Ricerca (MUR) Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) 2020 project RIPER-Resilient Artificial Intelligence-Based Self-Programming and Strategic Reasoning.

ABSTRACT The high volume of emergency room patients often necessitates head CT examinations to rule out ischemic, hemorrhagic, or other organic pathologies. A system that enhances the diagnostic efficacy of head CT imaging in emergency settings through structured reporting would significantly improve clinical decision making. Currently, no AI solutions address this need. Thus, our research aims to develop an automatic radiology reporting system by directly analyzing brain anomalies in head CT data. We propose a multi-branch CNN-LSTM fusion network-driven system for enhanced radiology reporting in emergency settings. We preprocessed head CT scans by resizing all slices, selecting those with significant variability, and applying PCA to retain 95% of the original data variance, ultimately saving the most representative five slices for each scan. We linked the reports to their respective slice IDs, divided them into individual captions, and preprocessed each. We performed an 80-20 split of the dataset for ten times, with 15% of the training set used for validation. Our model utilizes a pretrained VGG16, processing groups of five slices simultaneously, and features multiple end-to-end LSTM branches, each specialized in predicting one caption, subsequently combined to form the ordered reports after a BERT-based semantic evaluation. Our system demonstrates effectiveness and stability, with the postprocessing stage refining the syntax of the generated descriptions. However, there remains an opportunity to empower the evaluation framework to more accurately assess the clinical relevance of the automatically-written reports. Part of future work will include transitioning to 3D and developing an improved version based on vision-language models.

INDEX TERMS Convolutional neural network, emergency room, head computed tomography, language model, long short-term memory, radiology reporting.

Clinical and Translational Impact Statement—Our system improves clinical decision making by automating radiology reporting for emergency head CTs, enhancing diagnostic accuracy, reducing cognitive biases, and providing timely support for integration in hectic clinical settings.

I. INTRODUCTION

THE number of patients who access an emergency room is typically very significant and, at the same time, a head Computed Tomography (CT) examination is often required. The demand for head CT scans in emergency settings is driven by the need to quickly and accurately diagnose critical conditions such as strokes, traumatic brain injuries, and other neurological emergencies. However, this diagnostic process is not without challenges. The high costs associated with CT imaging include those related to the patient's extended stay in the hospital. Prolonged observation periods increase the risk of hospital-acquired conditions, including

© 2025 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

delirium and other complications, particularly in older adults.

Radiologists examine CT scans to diagnose the cause of patient's symptoms, monitor treatment effects, screen for various illnesses, and write radiological reports [1]. However, the process of interpreting CT scans is time consuming. A radiologist spends 5 to 20 minutes to interpret and manually describe the CT findings, further delaying patient's stay in the hospital [2]. In addition, this process is highly dependent on the individual experience. Radiologists rely on cognitive heuristics when making diagnostic decisions, such as representativeness, anchoring, and availability [3]. Representativeness refers to the tendency to assess the probability of a clinical condition based on how much the patient's symptoms resemble typical cases of that condition. Anchoring involves relying on the initial piece of information (i.e., the anchor), whereas availability is the tendency to judge the likelihood of events based on how easily examples come to mind. These heuristics can sometimes lead to diagnostic errors, especially in high-pressure emergency environments. Automatic radiology reporting represents a crucial advancement in reducing radiologists' workload and improving diagnostic accuracy [2]. Deep Learning (DL) models present challenges due to their non-transparent decision-making processes. This lack of interpretability can hinder trust and adoption in clinical settings. Teaching machines to produce automatic, human-readable reports is a promising approach to address this issue, as it can enhance the transparency of DL models by providing a clear rationale behind their decisions [4]. A well-designed decision-support system for radiology reporting can bridge the gap between raw data analysis and clinical decision making, offering a more intuitive understanding of the Artificial Intelligence (AI)'s findings and recommendations. Therefore, exploring the development of automatic radiology reporting systems is essential not only for improving the interpretability of DL models but also for ensuring that these systems can effectively support radiologists in making accurate, consistent, and unbiased diagnoses.

In this context, an automatic radiology reporting system designed to enhance the diagnostic efficacy of head CT imaging in emergency settings by harmonizing the report structure could be highly beneficial to the clinical decision-making process. Such a system could standardize the interpretation of head CT scans, reduce variability in diagnoses, and ensure that critical findings are consistently reported. In addition, creating multi-purpose reporting systems for radiologists that can detect several diseases simultaneously is still a challenge as, up to date, there are no solutions based on AI specifically designed to fulfill this task. Thus, the motivation behind our research is to create an automatic radiology reporting system that directly analyzes anomalies in brain structures, such as calcifications, hemorrhages, and other deviations from brain anatomies considered unremarkable, to provide radiologists with a second opinion and help reduce cognitive biases.

In this paper, we propose a multi-branch Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) fusion network-driven system with a Bidirectional Encoder Representations from Transformers (BERT)-based semantic evaluator for enhanced radiology reporting in emergency room head CT scans, whose workflow is illustrated in Fig. 1, articulating the objective into four key contributions:

- A dual-input preprocessing pipeline that optimizes both imaging and textual data, including a strategy to select the most representative slices of each scan when a slice-by-slice labeling is unavailable or not possible to perform;
- A multi-branch CNN-LSTM fusion network that integrates imaging features with multiple parallel LSTM branches, each receiving distinct inputs and producing a set of outputs, subsequently combined to form the ordered reports, thus leading to a stronger radiological data representation;
- A BERT-based evaluator that semantically analyzes and selects the best among the predictions for each section of the report, ensuring it appears properly structured;
- A rule-based postprocessing stage that refines the syntax of the automatically-written reports, making them more in line with how radiologists write.

It is also noteworthy to underline that this system is the first to work on head CT scans of emergency room patients, thus considering the extreme heterogeneity of cases that it has to deal with and its potential usefulness as a supportive tool in such a hectic clinical context.

For promoting transparency and reproducibility, we release the source code of the proposed system in a GitHub repository accessible at https://github.com/S3l11/HeadCTRadRepo. To safeguard privacy, raw data are not openly accessible. Access to anonymized textual data and preprocessed imaging data, represented as selectively-reduced and compressed versions of the original head CT scans, is subject to approval. Requests must be submitted in writing and include a declaration of non-profit use.

II. LITERATURE REVIEW

The initial focus of DL in radiology was on employing CNNs due to their effectiveness in processing complex imaging data [5]. CNNs excel at automatically amalgamating low-level features into high-level representations through successive layers. However, creating reports required the integration of a sequential model capable of handling textual data, leading to the adoption of Recurrent Neural Networks (RNNs) [6]. CNNs encode images into feature vectors, which RNNs then decode into textual descriptions, thus providing a more comprehensive understanding of radiological data. A notable example of this approach is the work by Shin et al. [7], the first that utilized CNNs to identify regions of interest in chest X-ray images from the Indiana University chest X-ray (IU X-ray) dataset and RNNs to

IEEE Journal of Translational Engineering in Health and Medicine

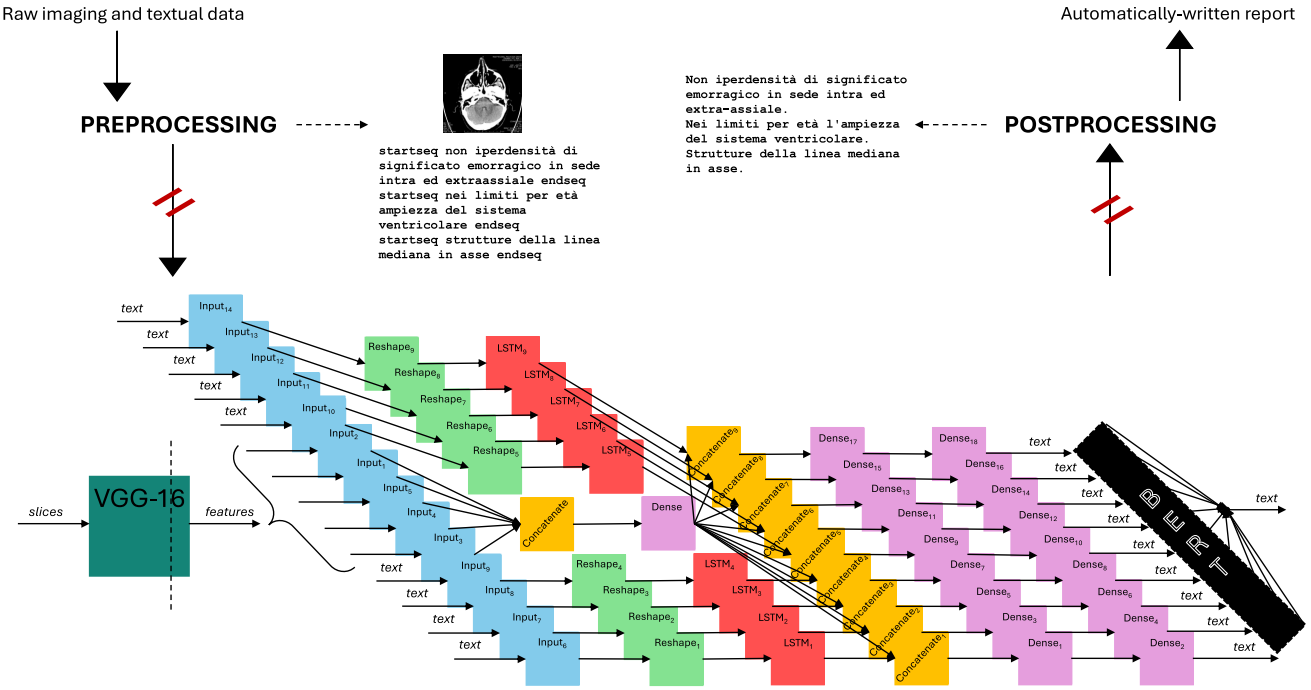


FIGURE 1. Workflow of the multi-branch CNN-LSTM fusion network-driven system. The process begins with a preprocessing pipeline followed by the extraction of relevant features using a pretrained VGG16. These features are processed through multiple parallel LSTM branches trained from scratch, each responsible for generating distinct captions corresponding to different aspects of the radiological findings. The BERT-based semantic evaluator analyzes and selects the best individual captions, which are then combined to form a complete, ordered report. A final postprocessing stage refines the syntax of the generated text, ensuring structural coherence. The red bars (//) indicate missing passages, excluded from this exemplification to ensure its clear reading.

produce corresponding descriptions (e.g., location, severity, and affected organs).

However, RNNs face challenges in retaining information over extended sequences, as their layers apply equal weights throughout. Additionally, training RNNs with the backpropagation algorithm often results in gradients that either explode or vanish, necessitating variations of the RNN architecture. The most prominent extension is the LSTM network, which incorporates memory blocks to preserve the network's temporal state and gates to regulate the flow of information [8]. Within a LSTM hidden unit, each sequence is processed in its entirety, with information stored in a memory cell. The memory cell itself determines what information to retain and when to allow its reading or updating through three gates, which operate as needed: the input gate transfers new information into the memory cell, the forget gate selectively discards irrelevant data, and the output gate enables the storage of important information. Specifically, the output h_t at time point t is regulated by (1), where i_t, f_t, o_t, and c_t are the activation vectors of the three gates and of the memory cell at time point t, σ is the sigmoid activation function, *tanh* is the hyperbolic tangent activation function, x_t denotes the current input, b denotes the bias of the memory cell and of each gate, and W are the diagonal weight matrices:

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i),\\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f),\\
c_t &= f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c),\\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o),\\
h_t &= o_t tanh(c_t).
\end{aligned} \tag{1}
$$

The first work that introduced a hierarchical LSTM to produce reports by capturing long-range semantics was by Jing et al. [9]. Although this model achieved outstanding results on the IU X-ray dataset, the predictions contained repeated sentences due to a lack of contextual coherence in the hierarchy. More recently, Xue et al. [10] created sentences using the same dataset by employing CNN and LSTM in a recurrent way. Similarly, Li et al. [11] proposed a system that first annotated the X-ray image via classifying and localizing common thoracic diseases and then created sentences from an attentive LSTM. However, the generated text lacked some abnormal descriptions and included sentences that differed from those in the training set. Recent advancements have significantly improved the quality and relevance of the generated text as the work by Sirshar et al. [12], who proposed an attention-based model that combined a CNN encoder with a LSTM decoder enhanced by an attention mechanism for sequence generation using the IU X-ray and Medical Information Mart for Intensive Care Chest X-Ray (MIMIC-CXR) datasets.

Despite the advancements of CNNs used in conjunction with recurrent modules, they encounter significant limitations in producing reports that are sufficiently diverse

and contextually accurate, both critical for clinical acceptance and utility. Their rigid structure frequently results in repetitive and less human-like descriptions [13]. Moreover, some of the datasets used to train these models often lack the realism needed for effective clinical application. Publicly available datasets like Radiology Objects in COntext (ROCO) [14] and ImageCLEF [15], which contain radiological images extracted from articles on PubMed Central and similar databases, do not fully reflect clinical environments. The associated captions are typically brief and lack the detail of true radiological reports, thus limiting the ability to capture the nuances of radiology reporting. Despite openly-accessible datasets like IU X-ray [16], ChestX-ray14 [17], CheXpert [18], and MIMIC-CXR [19] offer more realistic data, their focus on X-ray images restricts their applicability across various imaging modalities.

To our knowledge, only two works address CT data in this context. Aswiga et al. [20] developed a radiology reporting model that utilized a deep recurrent architecture, integrating a multi-level transfer learning framework with a LSTM. This model used image features, extracted from a private dataset of 150 breast and thorax CT images, and word vectors as inputs to produce captions. However, their approach involves a pure concatenation of captions rather than producing a full report. On the other hand, Kim et al. [21] proposed a CT scan-based captioning model that combined a CNN with a language model, specifically distilGPT2. Their study focused on normal and intracerebral hemorrhage head CT scans along with the corresponding reports. However, it is limited to the prediction of individual captions. In both cases, the CT images and associated reports are not publicly accessible, and the generated descriptions are neither fully comprehensive nor properly structured.

Our research aims to address these challenges by improving the diversity and contextual relevance of the automatically-written reports, thereby enhancing their suitability to support the clinical decision-making process.

III. DATA AND METHODOLOGY

A. IMAGING AND TEXTUAL DATA COLLECTION

We gathered data from the Synapse Teaching File of the Azienda Provinciale per i Servizi Sanitari of the Autonomous Province of Trento, Italy, comprising a total of 500 patients who visited the emergency room in the first half of 2022, from January to June. For each patient, we collected one head CT scan performed without intravenous contrast administration. The inclusion criteria encompassed both males and females over the age of 18, regardless of neurological pathology or injury status. We included not only cases with or without pathologies, but also cases with deviations from brain anatomies considered unremarkable. These deviations may not necessarily be linked to a pathological condition. Additionally, a single patient may present with multiple clinical conditions, mirroring real-world scenarios. Table 1 categorizes the included conditions, such as the presence of blood

TABLE 1. Occurrence, as %, and presence, as True/False, of the main clinical conditions included in the dataset. In line with the standard structure of the radiological reports, which are inherently comprehensive and include both positive (i.e., presence) and negative (i.e., explicit absence or negation) clinical findings, this stratification represents a holistic grouping.

Clinical condition	Occurrence	Presence in multiple reports
Hemorrhage	32.6	True
Calcifications	10	True
Ischemia	19.8	True
Edema	13.4	True
Other clinical findings	42.6	True

(e.g., hemorrhage), calcifications, ischemia, edema, and other findings. When multiple scans were available for a patient, we selected the earliest one to prevent intra-patient bias and potential information leakage. When also a head CT scan with intravenous contrast administration was available for a patient, we discarded it. Each scan consists of a variable number of slices, ranging from a minimum of 13 to a maximum of 307, with resolutions between 512 pixels $\times$ 512 pixels and 512 pixels $\times$ 559 pixels. During the collection phase, we anonymized all scans by removing any personal information (i.e., name, surname, date of bird, and sex) and the name of the hospital where the exam was undertaken, whereas retaining relevant information such as the field of view and compression rate. We stored the anonymized slices as JPEG files in zipped folders numbered in ascending order, one for each scan. Alongside imaging data, we also collected associated reports written in Italian. We anonymized these reports by removing any personal information and saved them as TXT files, one for each scan.

We performed both imaging and textual data collection in accordance with the General Data Protection Regulation-compliant University of Trento's guidelines and the European AI Act, following the ethical principles of the Helsinki Declaration and all applicable national laws governing observational feasibility studies.

B. PREPROCESSING STRATEGY

We undertook both imaging and textual data preprocessing before feeding them to our multi-branch CNN-LSTM fusion network.

1) IMAGING DATA PREPROCESSING

We extracted all slices from each zipped head CT scan, subsequently opening, converting them to gray scale, and resizing each to a uniform dimension of 224 pixels $\times$ 224 pixels. This standardization was crucial for ensuring consistency across the dataset. Throughout these initial steps, we meticulously managed arrays containing the slices, their resized counterparts, and their corresponding indices, safeguarding data integrity. After resizing, we analyzed the variance in pixel values across all slices, selecting those exhibiting significant variability, deemed to contain the most diagnostically-relevant features. Before applying Principal

Component Analysis (PCA) to the selected slices, we normalized data by subtracting the mean and dividing by the standard deviation to reduce variations in brightness due to diverse scanning parameters, addressing PCA's sensitivity to initial variable variances. We configured the PCA to retain 95% of the original data variance, aiming to preserve the most informative features while minimizing redundancy [22]. This involved calculating the covariance matrix, extracting its eigenvalues and eigenvectors, and selecting principal components that cumulatively accounted for the designated percentage of variance. This selective reduction not only streamlined data dimensionality but also enhanced computational efficiency by focusing on the most salient features [23]. We systematically saved the best five slices (i.e., those most characteristic based on their variance contributions and selected for striking the best balance between performance and computational efficiency, for a totality of 2500 slices) in a structured directory designed for optimal retrieval. We saved data in JPEG format for the slices to facilitate visual inspection, and in CSV format for the features, which included transformed features and their indices, ensuring comprehensive data documentation. Finally, for each patient's folder in the PCA-reduced dataset, we automatically checked all JPEG files and the associated TXT file. Each slice was then copied to an output directory with a new, systematic naming convention that incorporated the CT scan name to which the slices belong as the first four digits of the slice ID (i.e., XXXX_SliceYYYY). We paired each report with the slice IDs and recorded these pairings in a list, which served as a comprehensive index, mapping each preprocessed group of five slices to its corresponding report.

2) TEXTUAL DATA PREPROCESSING

We constructed a dictionary where each slice ID was mapped to a list of captions. These captions were extracted by reading through the corresponding file line by line. For each line, we identified the slice ID and its associated report, splitting the report into individual captions at each period, provided the resulting text was not just white space. The split of the report into individual captions served to reduce the complexity of the decoding process. By fragmenting the text into smaller, independent sentences, the decoding operation becomes more manageable and modular, ensuring a more efficient process, thus optimizing the system's ability to handle and interpret the information effectively, even when working with a relatively small dataset. Once the captions were loaded into our dictionary, we computed the maximum number (N_{max}) of captions per slice ID by finding the length of the longest list of captions in the dictionary to standardize input sequences. We than embarked on a more detailed cleaning process. We prepared a set of rules to eliminate punctuation, a necessary step to reduce noise in textual data. For each caption, we broke down the text into individual words, converted these words to lowercase to maintain consistency, and stripped away punctuation. Additionally, we removed any tokens that were a single character in length, as they generally do not contribute meaningful information. After cleansing the tokens, we rejoined them into a single string, book-ending each caption with markers (i.e., "startseq" and "endseq") to indicate the start and end of a caption, thus facilitating the model's learning of caption structure, as exemplified in Fig. 2. The cleaned captions were then stored back into a new dictionary, keyed by the original slice IDs. To ensure our imaging data corresponded with the preprocessed captions, we filtered out any slices whose IDs were not present in the new dictionary. This alignment ensured each slice to be associated with its correct, cleaned caption.

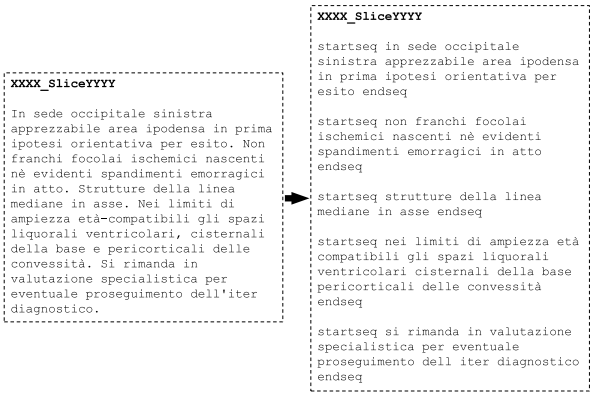


FIGURE 2. Result of textual data preprocessing (on the right), starting from one radiological report (on the left).

C. PREPROCESSED DATA HANDLING

To ensure the slices were grouped and split appropriately, we organized them based on the first four digits of their IDs, guaranteeing that each PCA-reduced group contained exactly five slices. Captions for each slice were tokenized into integer sequences, padded to a uniform length. These sequences were then used to build input and output data for the model, with sequences padded and outputs encoded categorically.

We conducted a randomized 80-20 split of the dataset, allocating 80% for training and 20% for testing. Hence, out of the 2500 total slices, we used 2000 for training and 500 for testing, and we designated 15% of the training set (i.e., 300 slices) as the validation set. We repeated this randomized 80-20 split ten times to ensure that the results were not biased by a potentially favorable data split.

D. MODEL ARCHITECTURAL DETAILS

We designed our multi-branch CNN-LSTM fusion network with the five slice feature inputs and the nine caption inputs, producing nine separate textual outputs. The use of multiple parallel LSTMs allowed the model to capture temporal characteristics and long-term dependencies within the caption sequences, while preserving the relative position of each caption in the report. For instance, the first LSTM processed all captions appearing in the first position, the second LSTM handled captions in the second position, and so on, covering

TABLE 2. Model architecture. $Input_1$ to $Input_5$ are feature vectors extracted beforehand using a pretrained VGG16 and concatenated into a single vector of size 20480 (4096 × 5). The concatenated vector is then passed through a Dense layer. $Input_6$ to $Input_{14}$ represent sequences of text, each with a maximum length of 82. Each caption input is then reshaped to be compatible with and processed through a LSTM ($LSTM_1$ to $LSTM_9$) layer, which outputs a feature vector of size 256. The combined slice features are concatenated with the features generated by the LSTM layers and each concatenated vector is then passed through a Dense layer, reducing its size to 256. This is followed by another Dense layer with a Softmax activation function, producing the final output.

Layer	Output shape	Number of parameters	Connected to layer
$Input_1$	(None, 4096)	0	[]
$Input_2$	(None, 4096)	0	[]
$Input_3$	(None, 4096)	0	[]
$Input_4$	(None, 4096)	0	[]
$Input_5$	(None, 4096)	0	[]
$Input_6$	(None, 82)	0	[]
$Input_7$	(None, 82)	0	[]
$Input_8$	(None, 82)	0	[]
$Input_9$	(None, 82)	0	[]
$Input_{10}$	(None, 82)	0	[]
$Input_{11}$	(None, 82)	0	[]
$Input_{12}$	(None, 82)	0	[]
$Input_{13}$	(None, 82)	0	[]
$Input_{14}$	(None, 82)	0	[]
Concatenate	(None, 20480)	0	[$Input_1$, $Input_2$, $Input_3$, $Input_4$, $Input_5$]
$Reshape_1$	(None, 82, 1)	0	[$Input_6$]
$Reshape_2$	(None, 82, 1)	0	[$Input_7$]
$Reshape_3$	(None, 82, 1)	0	[$Input_8$]
$Reshape_4$	(None, 82, 1)	0	[$Input_9$]
$Reshape_5$	(None, 82, 1)	0	[$Input_{10}$]
$Reshape_6$	(None, 82, 1)	0	[$Input_{11}$]
$Reshape_7$	(None, 82, 1)	0	[$Input_{12}$]
$Reshape_8$	(None, 82, 1)	0	[$Input_{13}$]
$Reshape_9$	(None, 82, 1)	0	[$Input_{14}$]
Dense	(None, 4096)	83890176	[Concatenate]
$LSTM_1$	(None, 256)	264192	[$Reshape_1$]
$LSTM_2$	(None, 256)	264192	[$Reshape_2$]
$LSTM_3$	(None, 256)	264192	[$Reshape_3$]
$LSTM_4$	(None, 256)	264192	[$Reshape_4$]
$LSTM_5$	(None, 256)	264192	[$Reshape_5$]
$LSTM_6$	(None, 256)	264192	[$Reshape_6$]
$LSTM_7$	(None, 256)	264192	[$Reshape_7$]
$LSTM_8$	(None, 256)	264192	[$Reshape_8$]
$LSTM_9$	(None, 256)	264192	[$Reshape_9$]
$Concatenate_1$	(None, 4352)	0	[Dense, $LSTM_1$]
$Concatenate_2$	(None, 4352)	0	[Dense, $LSTM_2$]
$Concatenate_3$	(None, 4352)	0	[Dense, $LSTM_3$]
$Concatenate_4$	(None, 4352)	0	[Dense, $LSTM_4$]
$Concatenate_5$	(None, 4352)	0	[Dense, $LSTM_5$]
$Concatenate_6$	(None, 4352)	0	[Dense, $LSTM_6$]
$Concatenate_7$	(None, 4352)	0	[Dense, $LSTM_7$]
$Concatenate_8$	(None, 4352)	0	[Dense, $LSTM_8$]
$Concatenate_9$	(None, 4352)	0	[Dense, $LSTM_9$]
$Dense_1$	(None, 256)	1114368	[$Concatenate_1$]
$Dense_3$	(None, 256)	1114368	[$Concatenate_2$]
$Dense_5$	(None, 256)	1114368	[$Concatenate_3$]
$Dense_7$	(None, 256)	1114368	[$Concatenate_4$]
$Dense_9$	(None, 256)	1114368	[$Concatenate_5$]
$Dense_{11}$	(None, 256)	1114368	[$Concatenate_6$]
$Dense_{13}$	(None, 256)	1114368	[$Concatenate_7$]
$Dense_{15}$	(None, 256)	1114368	[$Concatenate_8$]
$Dense_{17}$	(None, 256)	1114368	[$Concatenate_9$]
$Dense_2$	(None, 1153)	296321	[$Dense_1$]
$Dense_4$	(None, 1153)	296321	[$Dense_3$]
$Dense_6$	(None, 1153)	296321	[$Dense_5$]
$Dense_8$	(None, 1153)	296321	[$Dense_7$]
$Dense_{10}$	(None, 1153)	296321	[$Dense_9$]
$Dense_{12}$	(None, 1153)	296321	[$Dense_{11}$]
$Dense_{14}$	(None, 1153)	296321	[$Dense_{13}$]
$Dense_{16}$	(None, 1153)	296321	[$Dense_{15}$]
$Dense_{18}$	(None, 1153)	296321	[$Dense_{17}$]
		Trainable: 98964105 **Non-trainable**: 0	

all nine positions across different reports. This model architecture, displayed in Table 2, allowed us to integrate visual and textual data effectively, leveraging the strengths of both types of information.

We began by extracting essential features from the slices using a pretrained VGG16 [24]. This network, consisting of 16 layers with small receptive fields (3 × 3 filters), was restructured to output the activations from the second-to-last layer, providing a 4096-dimensional feature vector for each slice. Specifically, we defined five inputs for the slice features, each with a shape of (4096,). These features were concatenated into a single vector of 20480 dimensions

IEEE Journal of Translational Engineering in Health and Medicine

(5 slices × 4096 dimensions) and passed through a Dense layer with 4096 neurons and a Rectified Linear Unit (ReLU) activation function to combine the grouped slice features. We also defined nine inputs for the captions, each with a shape of (82,), representing the tokenized and padded sequences of words. These inputs were reshaped to (82, 1) to be processed by nine parallel end-to-end LSTM layers with 256 units each, particularly useful for sequence prediction problems. Each LSTM processed captions corresponding to a specific position in the sequence, and the outputs of these LSTM layers were then concatenated with the processed slice features, resulting in combined feature vectors of 4352 dimensions. Each combined vector was passed through a Dense layer with 256 neurons and a ReLU activation function to make the model take into account non-linear interactions. The outputs of these Dense layers were further processed through final output layers, each producing a probability distribution over the vocabulary words using a Softmax activation function to assign probabilities to each class by producing real values between 0 and 1, with sum equal to 1.

E. MODEL FINAL CONFIGURATION

We customized the final architecture of our multi-branch CNN-LSTM fusion network after two sets of preliminary experiments. In the first set of experiments, we tested state-of-the-art CNNs as backbones (i.e., feature extractors) to find the best performing for the addressed task. In the second set of experiments, we searched for the most suitable values for the hyperparameters, specifically the number of epochs, learning rate, and batch size. For all experiments, we chose the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 2. Each time a new best-performing value for any hyperparameter was found, it replaced the one proposed in this base configuration.

1) OPTIMAL BACKBONE SELECTION

We tested a well-established CNN architecture such as VGG16 as baseline, known for its strong performance in image recognition tasks due to its architecture. We also considered ResNet50V2 [25], Xception [26], and Inception-ResNetV2 [27] due to their competitive performance on ImageNet dataset. The comparison showed that VGG16 was the best performing (i.e., the one with the lowest validation loss) to extract visual features. Its pretrained weights provided robust generalization capabilities, allowing it to be used as pure feature extractor without fine-tuning. This also allowed resources to be allocated to training the multiple parallel LSTMs from scratch. Thus, following experimentation made use of VGG16 as backbone for the final structure of our model.

2) OPTIMAL HYPERPARAMETER SELECTION

Using the best-performing backbone from the previous set of experiments, the choice of the optimal hyperparameter combination was driven by the Bayesian optimization algorithm, recognized for its effectiveness in maximizing model performance [28]. We investigated all possible combinations among the following values, chosen for their computational efficiency in preliminary evaluations:

- Number of epochs: [50, 100, 150];
- Learning rate: [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05];
- Batch size: [2, 4, 8, 16, 32, 64].

We selected the hyperparameter combination that led to the lowest validation loss. Thus, following experimentation made use of 100 epochs, 0.001 as learning rate, and 32 as batch size.

F. COMPUTATIONAL ENVIRONMENT AND TRAINING STRATEGY

We set up the computational environment using the Keras library built on a TensorFlow backend (version 2.15.0) and one NVIDIA A100 GPU with 40 GB of RAM for accelerated computing.

We trained our multi-branch CNN-LSTM fusion network for 100 epochs, with a learning rate set to 0.001 and a batch size of 32. We confirmed Adam as the optimizer, recognizing its effectiveness in efficiently minimizing the loss function during training. We used categorical cross-entropy as loss function and incorporated an early stopping callback with a patience of 8 to further mitigate overfitting. Training the model from scratch took approximately 6 hours. We then saved the model weights that resulted in the lowest validation loss and used these weights to evaluate model performance on the test set.

G. REPORTING PROCESS AND EVALUATION

We first initialized nine tokenizers, each fitted on the textual data from the caption dictionary, to build a vocabulary that mapped words to unique integers based on their frequency in the text. Next, we grouped the test slices by their prefixes, corresponding to the first four characters of each slice ID, which allowed us to handle all slices from a single CT scan together and ensure consistent comparisons across related slices. The model then produced captions for each slice in the group, generating a set of captions for the nine positions managed by the tokenizers. If fewer than nine captions were produced, we added empty strings for consistency. The reporting process began with a predefined start token and proceeded iteratively, with each step generating the next word based on the previously generated text and the visual features extracted from each slice. The model selected the next word from a probability distribution over the vocabulary, influenced by a temperature parameter that controlled the randomness of predictions. Lower temperatures led to more predictable, conservative word choices, whereas higher temperatures introduced more variability and creativity. This process continued until the model predicted an end token or reached the maximum predefined length for each description.

For each produced caption, we calculated BiLingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and METEOR scores. These metrics automatically compute accuracy scores by comparing the predicted reports with the radiologists' written descriptions. BLEU scores were calculated for different n-gram combinations (BLEU-1 through BLEU-4) as in (2), (3), (4), and (5), providing a detailed measure of precision in the generated text [29], where BP is the brevity penalty and p_n is the n-gram precision:

$$BLEU\text{-}1 = BP p_1, \tag{2}$$

$$BLEU\text{-}2 = BP \sqrt{p_1 p_2}, \tag{3}$$

$$BLEU\text{-}3 = BP \sqrt[3]{p_1 p_2 p_3}, \tag{4}$$

$$BLEU\text{-}4 = BP \sqrt[4]{p_1 p_2 p_3 p_4}. \tag{5}$$

ROUGE scores focus on recall, assessing the overlap between generated and reference texts [30], with ROUGE-1 (6), ROUGE-2 (7), and ROUGE-L (8) being used, where LCS is the length of the longest common subsequence, as ROUGE-L specifically measures the longest common subsequences between two sentences:

$$ROUGE\text{-}1 = \frac{\sum_{1\text{-}gram \in Reference} Count_{match}(1\text{-}gram)}{\sum_{1\text{-}gram \in Reference} Count(1\text{-}gram)}, \tag{6}$$

$$ROUGE\text{-}2 = \frac{\sum_{2\text{-}gram \in Reference} Count_{match}(2\text{-}gram)}{\sum_{2\text{-}gram \in Reference} Count(2\text{-}gram)}, \tag{7}$$

$$ROUGE\text{-}L = \frac{LCS(Candidate, Reference)}{Length(Reference)}. \tag{8}$$

METEOR score, computed as in (9), offers a more balanced view by considering both precision and recall alongside additional linguistic factors [31], addressing some of BLEU's limitations by incorporating synonym matching through WordNet, where $F_{mean} = \frac{10PR}{R+9P}$ and $Penalty = \gamma \left(\frac{chunks}{matches}\right)^{\theta}$, being P precision, R recall, and γ and θ tuning parameters:

$$METEOR = F_{mean}(1 - Penalty). \tag{9}$$

BLEU emphasizes precision, whereas METEOR and ROUGE are recall-based metrics, complementing each other in evaluating the quality, accuracy, and robustness of the generated text [32].

Once the N_{max} captions were produced for each slice of the PCA-reduced group of five slices belonging to the test set, they were evaluated using BERT [33] as pure semantic evaluator, making possible to calculate the perplexity score for each caption. We chose BERT to assess the likelihood of the predicted captions, with lower perplexity scores indicating higher semantic coherence and more natural language structure. For each position, the captions were scored based on this perplexity measure, and the caption with the lowest perplexity (i.e., the highest semantic quality) was selected as the best option for that position. If no valid caption was available for a specific position, a fallback mechanism ensured that the first prediction generated by the corresponding predictor for that position was selected instead. This process ensured that, even in the absence of high-quality options, a caption was always chosen. We then compiled the best captions for all positions in order to form the complete, ordered report for each group of five slices. The resulting best predictions for each CT scan were compared with the respective ground truths by calculating BLEU-1 to BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR scores. Fig. 3 displays the reporting process, using as example a PCA-reduced group of five slices from a test head CT scan. To summarize overall performance, we finally calculated the global averages of each metric across the entire test set, providing a comprehensive view of the accuracy and consistency of the model.

H. POSTPROCESSING STRATEGY

We undertook textual data postprocessing to refine the generated descriptions. Specifically, we made grammatical corrections without altering the semantics, ensuring that the content of the predictions remained unchanged.

We employed a rule-based approach, which included the use of regular expressions (i.e., regex) to match and correct common patterns, such as missing punctuation, incorrect capitalization, and inconsistent spacing. This procedure allowed us to systematically address typical formatting errors in the automatically-written reports, enhancing their readability. We then conducted a manual revision to make the refined reports adhere closely to the formatting standards of radiological ones. This additional step ensured that any errors missed by the rule-based approach were corrected while preserving the semantic integrity of the predictions.

I. ABLATION STUDY

We designed a baseline configuration to investigate the impact of treating the radiological report as a single cohesive unit, as opposed to decomposing it into individual caption sequences. In this configuration, slice features extracted from the VGG16 were passed through a single LSTM, which then produced the entire report as a monolithic text output.

The rationale behind this ablation study was to test whether simplifying the problem by removing the need to split reports and reducing the complexity of multiple parallel LSTMs could still support accurate radiology reporting. In this baseline, we also achieved the optimal trade-off between model performance and computational efficiency. By using this configuration as reference point, we aim to clearly demonstrate the performance gains achieved through problem decomposition and highlight the advantages of using multiple parallel LSTMs for processing independent captions.

J. ADDITIONAL STATISTICAL ANALYSIS

We applied the Friedman test to assess the presence of statistically significant differences among the test set combinations [34], setting the significance level (P) at 0.05. In the event that general differences were detected, we planned to conduct a post-hoc analysis using the Nemenyi test to

IEEE Journal of Translational Engineering in Health and Medicine

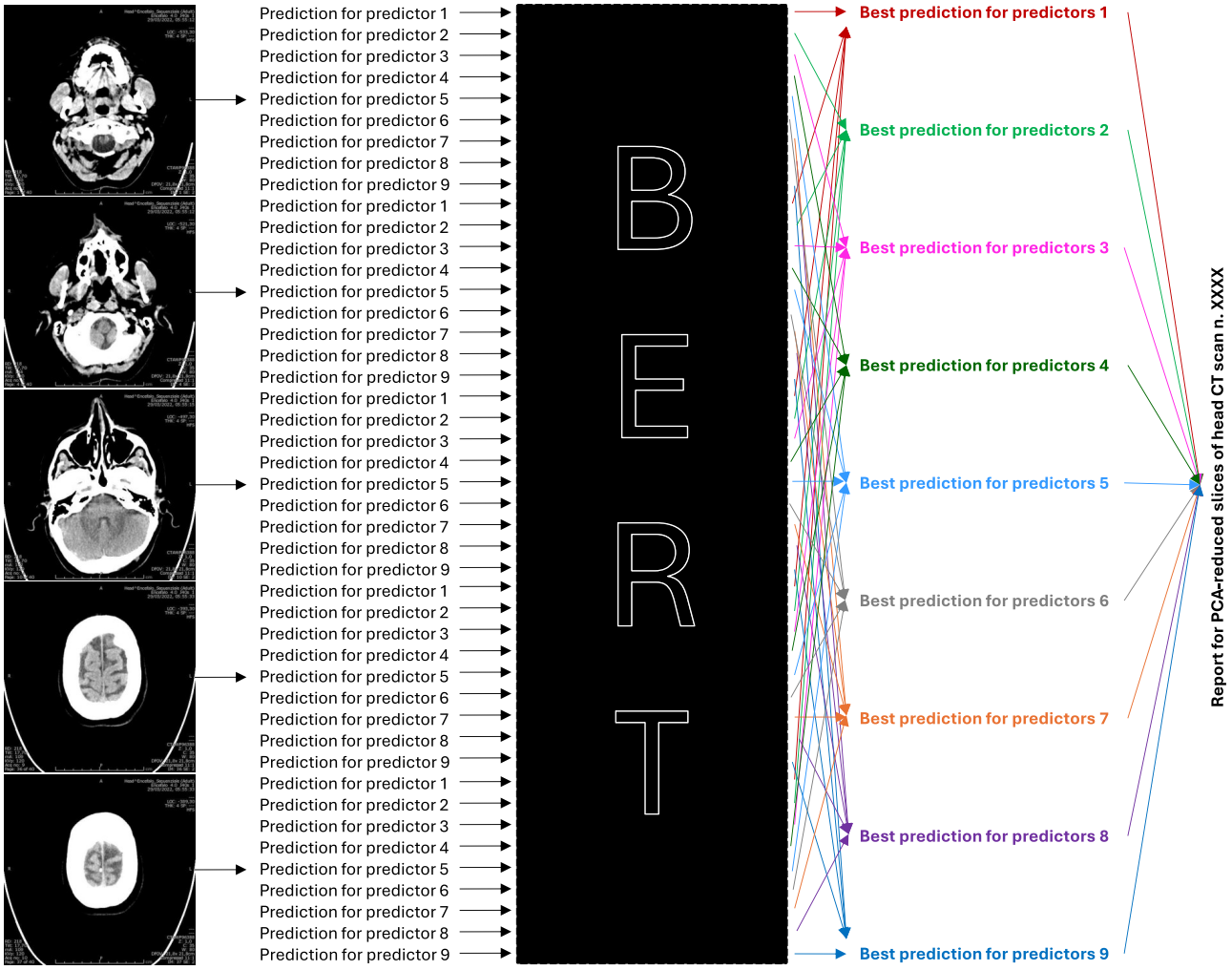


FIGURE 3. Automatic process of report creation for the PCA-reduced group of five slices of each test head CT scan. Nine (i.e., N_{max}) predictors generate distinct captions for each slice of the group. The predicted captions are then analyzed by the BERT semantic evaluator. For each position, the best predicted captions (i.e., those with the lowest perplexity score) are automatically selected and fused to compile the complete, ordered report for that specific scan.

identify any specific pairs of combinations with significant differences.

To compare the performance of the baseline configuration, the final model configuration, and the final model configuration with the added postprocessing stage, we employed the Wilcoxon signed-rank test [35]. This test was chosen to account for the paired nature of the data, as all configurations were evaluated on the same test set. We used this test to determine whether the differences in evaluation scores were statistically significant, with P again set at 0.05.

K. LANGUAGE IMPACT ASSESSMENT

We conducted a comprehensive evaluation of the final model configuration performance on radiological reports translated into English to ascertain whether the language of the input text impacted the accuracy and coherence of the generated descriptions.

The experimental setup involved the use of a verified translation pipeline to convert the original Italian reports into English while preserving semantic fidelity and consistency. The preprocessing pipeline remained unchanged, including textual data cleaning, tokenization, and the split of the reports into individual captions. We utilized the same multi-branch CNN-LSTM fusion network architecture, along with identical hyperparameters. To maintain consistency, we performed the same data division protocol (80-20 for training and testing, with 15% of the training set used for validation). Each translated report was processed in alignment with its paired imaging data, ensuring compatibility with the PCA-reduced groups of five slices. The training and evaluation phases followed the same procedures, with performance assessed using BLEU, ROUGE, and METEOR metrics, along with semantic checks via BERT-based evaluation.

IV. RESULTS

In this section, we present the results in both graphical and tabular formats, and discuss them in section V.

TABLE 3. Evaluation scores for radiology reporting, expressed in %, over the ten different test set combinations. If present, * indicates statistically significant differences according to the Friedman test.

Test set combination	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
1	33.6	26.2	21	16.3	41.3	22.1	34.6	34.1
2	32.9	25.8	21.9	16.4	41.3	23.4	32.9	34.3
3	34	25.6	21.4	17.1	40.6	21.2	31	32.3
4	33.1	23.9	20.7	16.9	39.8	22.5	33.8	33.4
5	32.9	24.3	19.8	16.8	39	22.5	31.1	31.9
6	35.1	25.8	21.1	17.2	37.1	23.9	31	32.9
7	31.9	23.7	19	15.9	37.9	20.7	33.1	34.1
8	34.2	25.7	20.5	**17.3**	39.6	23.5	**34.6**	**34.5**
9	34.2	25.3	21.1	17	39.2	23.7	31.1	32.2
10	33.1	25.2	20.7	16	40	24.9	32.1	31.9

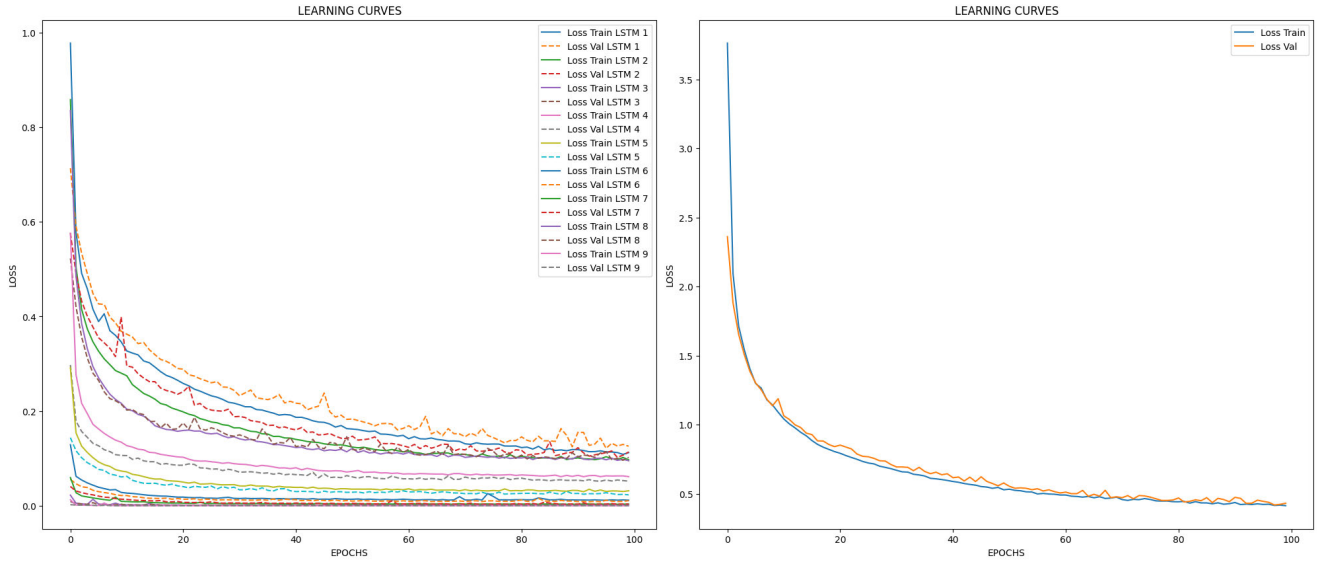


FIGURE 4. Individual (on the left) and overall (on the right) learning curves for the multi-branch CNN-LSTM fusion network *-ITA version-*.

Table 3 reports the evaluation scores for radiology reporting, including BLEU-1 to BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR, over the ten different test set combinations. These metrics are presented for the final model configuration only. Since the Friedman test revealed no statistically significant differences among the sets (i.e., $P > 0.05$), we chose to proceed with the combination that yielded the highest BLEU-4, ROUGE-L, and METEOR scores, precisely the 8^{th}. Fig. 4 illustrates the individual and overall learning curves for the multi-branch CNN-LSTM fusion network in the Italian version, whereas Fig. 5 illustrates the individual and overall learning curves for the multi-branch CNN-LSTM fusion network in the English version. Table 4 summarizes the evaluation scores for radiology reporting, including BLEU-1 to BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. These metrics are presented for the baseline configuration (i.e., single CNN-LSTM fusion network) in the Italian version, the final model configuration (i.e., multi-branch CNN-LSTM fusion network) in both Italian and English versions, and the final model configuration with the added postprocessing stage in both Italian and English versions, all computed on the same test set combination. The Wilcoxon signed-rank test revealed no statistically significant differences between the two final model configurations (Italian and English) and also between the final model configuration and the same configuration with the added postprocessing stage in both Italian and English versions, whereas all evaluation scores of the baseline model configuration are statistically lower with respect to those of the other configurations in the Italian version, except for ROUGE-1. Fig. 6 displays the descriptions generated in Italian as predicted by our model, alongside the postprocessed versions, compared with the reference reports for the PCA-reduced groups of five slices from three head CT scans belonging to the test set.

V. DISCUSSION

Clinical decision making is critical in the healthcare domain and errors in radiology reporting can lead to serious consequences. Therefore, improving the accuracy of reports by AI strategies is necessary, as there is still a significant gap in research in the related field. In light of this, we propose the first system for improved radiology reporting in emergency room head CT scans, consisting of a dual-input

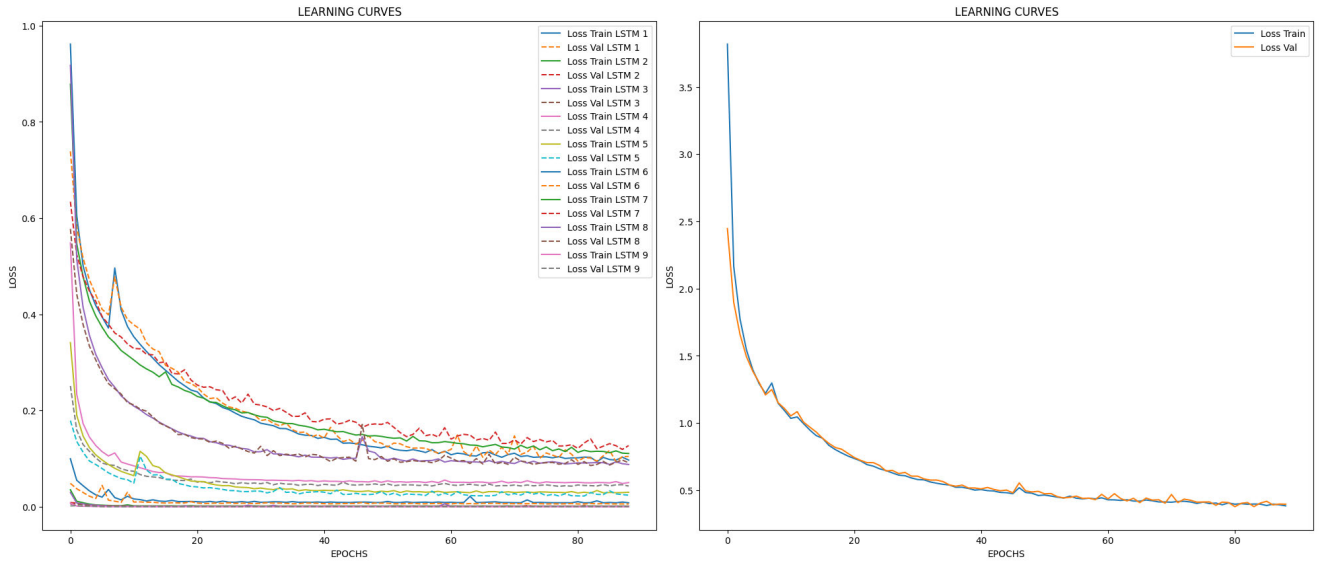


FIGURE 5. **Individual (on the left) and overall (on the right) learning curves for the multi-branch CNN-LSTM fusion network *-EN version-*.**

TABLE 4. **Evaluation scores for radiology reporting, expressed in %, on the test set. If present, * indicates statistically significant differences according to the Wilcoxon signed-rank test.**

Configuration	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Single CNN-LSTM fusion network (baseline)	27.9*	18.1*	12.6*	9.7*	36.8	16*	30.1*	28*
Multi-branch CNN-LSTM fusion network (final) *-ITA version-*	34.2	25.7	20.5	17.3	39.6	23.5	34.6	34.5
Multi-branch CNN-LSTM fusion network (final) *-EN version-*	33.2	25.9	19	16.8	37.4	22.9	32.7	32
Multi-branch CNN-LSTM fusion network (final) + postproc. *-ITA version-*	**34.4**	**26.2**	**22.1**	**19.4**	**39.7**	**24.9**	**35.6**	**37.7**
Multi-branch CNN-LSTM fusion network (final) + postproc. *-EN version-*	33.2	25.9	19.1	17.3	37.5	23.4	32.9	35.6

preprocessing pipeline, a multi-branch CNN-LSTM fusion network, a BERT-based semantic evaluator, and a rule-based postprocessing stage.

According to the achieved results, the absence of statistical significance in Table 3 suggests that no single test set combination is notably more favorable than others, indicating that our model generalizes well across different data combinations. This reflects stability and consistency in the learning process, regardless of the specific data split. The learning curves in Fig. 4 and Fig. 5 show a consistent decrease in loss for both the training and validation sets, implying steady model improvement. The close alignment between the training and validation loss further suggests that our model is not overfitting, as its performance remains comparable across both sets. Additionally, the curves converging to a similar value towards the end indicate that our multi-branch CNN-LSTM fusion network has likely reached its maximum learning potential for the given data. The architectural design of the model also helps mitigate overfitting risks. By combining features using a Dense layer with ReLU activation after the concatenation of slice and caption features, the network can focus on the most relevant interactions, reducing redundancy in the high-dimensional input space. Furthermore, the use of separate LSTMs for processing captions in specific positions ensures that each sequence is handled independently, minimizing the risk of unintended interactions between sequential elements. As evidenced in Table 4, the final model configuration achieves strong performance, with further improvements from the postprocessing stage, particularly in aligning reports with the structural conventions of the ground truth data. Compared to the baseline, these results highlight the clear benefits of problem decomposition and use of parallel LSTMs for handling different report sections independently. The model also maintained comparable levels of performance across both languages. Specifically, the BLEU-4, ROUGE-L, and METEOR scores for the English-translated reports are within a 3% margin of those obtained for the Italian ones, indicating that the model effectively generalized to the new linguistic context. This capability can be attributed to the parallel structure of the LSTM branches, trained to remain neutral to the language of the textual input. Additionally, the semantic evaluation stage, powered by BERT, further ensured language-agnostic assessment of the produced reports. This structured approach proves to be crucial for effective radiology reporting. However, the current evaluation metrics may not fully capture the quality of the predictions, particularly when the model's output closely matches the ground truth in content but diverges stylistically, leading to lower scores. For example, the predicted report for the first PCA-reduced group of five test

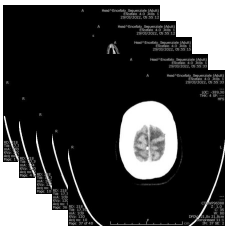
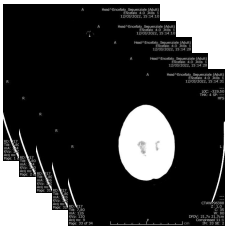
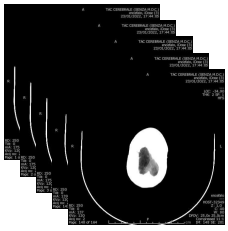
	PREDICTED REPORT	POSTPROCESSED REPORT	GROUND TRUTH	
	non sono attualmente apprezzabili segni tc di spandimenti ematici intra extraassiali nè di lesioni ischemiche recenti non alterazioni densitometriche focali acute intra ed extraassiali linea mediana in asse con sistema ventricolare in sede di regolare morfologia nella norma gli spazi subaracnoidei della convessità della base cranica cisternali	Non sono attualmente apprezzabili segni TC di spandimenti ematici intra ed extra-assiali nè di lesioni ischemiche recenti. Non alterazioni densitometriche focali acute intra ed extra-assiali. Linea mediana in asse con sistema ventricolare in sede, di regolare morfologia. Nella norma gli spazi subaracnoidei della convessità, della base cranica, cisternali.	Non spandimenti ematici in atto. Non alterazioni densitometriche focali acute intra ed extra-assiali. Linea mediana in asse con sistema ventricolare in sede, di regolare morfologia. Nella norma gli spazi subaracnoidei della convessità, della base cranica e cisternali.	**BLEU-1:** 72.3% **BLEU-2:** 69.8% **BLEU-3:** 68.2% **BLEU-4:** 66.9% **ROUGE-1:** 86.1% **ROUGE-2:** 76.5% **ROUGE-L:** 86.1% **METEOR:** 91.6%
	non iperdensità di significato emorragico in sede intra ed extraassiale non alterazioni densitometriche focali acute intra ed extraassiali non evidenti allo stato attuale alterazioni densitometriche focali carattere edemigeno acuto del parenchima encefalico linea mediana in asse con sistema ventricolare in sede di regolare morfologia	Non iperdensità di significato emorragico in sede intra ed extra-assiale. Non alterazioni densitometriche focali acute intra ed extra-assiali. Non evidenti allo stato attuale alterazioni densitometriche focali di carattere edemigeno acuto del parenchima encefalico. Linea mediana in asse con sistema ventricolare in sede, di regolare morfologia.	Non emorragie endocraniche di recente insorgenza. Non evidenti allo stato attuale alterazioni densitometriche focali a carattere edemigeno acuto del parenchima encefalico. Regolari le dimensioni del sistema ventricolare, delle cisterne della base e degli spazi liquorali periencefalici. Strutture della linea mediana in asse.	**BLEU-1:** 50% **BLEU-2:** 44.5% **BLEU-3:** 40.8% **BLEU-4:** 37.3% **ROUGE-1:** 60% **ROUGE-2:** 43.6% **ROUGE-L:** 54.3% **METEOR:** 52.3%
	spandimenti ematici alterazioni densitometriche focali carattere edemigeno acuto del parenchima encefalico	Spandimenti ematici. Alterazioni densitometriche focali di carattere edemigeno acuto del parenchima encefalico.	Grossolano spandimento emorragico a livello della sostanza bianca profonda emisferica sinistra con edema perilesionale ed estensione a livello degli spazi subaracnoidei fronto-parietali con emoventricolo. Compressione sul trigono del ventricolo omolaterale e modesto shift della linea mediana.	**BLEU-1:** 1.2% **BLEU-2:** 0.4% **BLEU-3:** 0.3% **BLEU-4:** 0.3% **ROUGE-1:** 0.5% **ROUGE-2:** 0.1% **ROUGE-L:** 0.5% **METEOR:** 1.6%

FIGURE 6. Generated descriptions as predicted by the multi-branch CNN-LSTM fusion network and postprocessed ones with the ground truths for the PCA-reduced groups of five slices of three head CT scans belonging to the test set.

slices (Fig. 6) demonstrates very high degree of alignment with the reference report in both content and style, indicating almost complete overlap. For the second PCA-reduced group of five test slices, the produced report shows strong content alignment with the reference report and maintains a good level of consistency in style. In contrast, the generated text for the third PCA-reduced group of five test slices exhibits good content alignment but poor stylistic alignment, despite identifying important clinical details. Specifically, the reference report indicates an hemorrhagic event in the left hemisphere, leading to significant swelling, extension of bleeding into other brain spaces, and a mild shift of the brain's central structures, necessitating immediate medical intervention. The predicted report notes localized swelling in brain tissue that requires medical attention, thus it effectively captures the key clinical problem even if lacking the detailed nuance found in the reference one. This highlights the importance of having individual scores for each generated description for a more granular assessment of its reliability but also the need for more refined metrics that better account for both content accuracy and stylistic coherence.

In the literature, some researchers have concentrated solely on extracting global features from radiological images, which often results in imprecise localization of anomalies that are typically confined to specific regions of the image [36], [37]. Additionally, many of the predicted reports are produced using RNNs, which are prone to the issue of gradient vanishing, particularly when dealing with long sentences [38]. Furthermore, several existing techniques struggle with generating words in the correct order [39] or crafting descriptions that are convincingly human-like [13], reducing their effectiveness in clinical practice. It is also important to note that no research group has yet tackled the challenge of making the radiology reporting task even more challenging by processing and analyzing data from emergency settings, where the range of cases is extremely broad and decisions need to be made within very limited timeframes. Compounding these issues is the fact that most available datasets are heavily focused on X-ray images and often include captions that are brief and lack the richness in detail characteristic of true radiological reports. Moreover, these datasets frequently fall short of the realism necessary for practical clinical application, as discussed in section II. The system proposed in this paper addresses and resolves all of these challenges, offering a more precise and clinically-relevant solution for radiology reporting. In fact, our multi-branch CNN-LSTM fusion network not only improves the localization of anomalies through more accurate feature extraction but also produces more coherent and human-like descriptions, thus enhancing its applicability in real-world clinical settings. Furthermore, training on diverse emergency-specific data ensures exposure to a wide spectrum of clinical conditions, thereby enhancing the system's robustness and generalizability.

The first limitation of the research presented in this paper lies in the absence of a label (i.e., clinical annotation) for each slice, due to the labor-intensive and time-consuming nature of

IEEE Journal of Translational Engineering in Health and Medicine

this task for radiologists. Additionally, the approach used is not fully 3D, as it processes and analyzes only the best five slices resulting from the PCA-based strategy for each head CT scan. This lowers the computational complexity, a critical requirement in emergency settings, but it may limit the ability of the model to capture the full spatial context and intricate details that are inherent in volumetric data [40]. Moreover, the multi-branch architecture is designed to process captions corresponding to specific sections of the report rather than explicitly targeting conditions like hemorrhage or ischemia. While this design ensures flexibility and coherence in the generated descriptions, it may limit diagnostic precision. Furthermore, despite the demonstrated effectiveness and stability of our system, particularly with the added postprocessing stage, there remains room for refinement in both model performance and evaluation framework to better assess the clinical relevance and medical accuracy of the automatically-written reports.

In future research, we plan to conduct a detailed stratification of the dataset's demographic characteristics to better align its representativeness with larger population distributions and further validate its applicability to diverse clinical scenarios. We will also refine our system by exploring alternative strategies, such as 3D approaches or resampling to uniform resolution and saving slices within a clinically-standardized 15-cm range centered on the brain, for a more robust spatial context analysis. We plan to enhance it further through the integration of vision-language models. These models have recently shown significant potential in health informatics, particularly for the automatic generation of reports by effectively integrating visual and textual data. However, they are still constrained by challenges related to the need for substantial computational resources and extensive domain-specific training data, which can impact their reliability and trustworthiness in clinical settings [41]. Ultimately, we will incorporate qualitative evaluations to prove that the PCA-based dimensionality reduction effectively retains diagnostically-relevant information, ensuring that the analyzed slices adequately represent the ground truth conditions. Qualitative metrics will also be employed to evaluate diagnostic accuracy by answering specific questions regarding the presence of errors or omissions in the produced reports (e.g., Are non-existent pathologies included? Are existing pathologies omitted?) and assessing stylistic coherence, thereby enhancing the overall robustness of our system.

VI. CONCLUSION

Automatic radiology reporting can significantly reduce the workload of report writing, but linking visual and textual knowledge from radiological data is still a task that has not been effectively solved. In the literature, researchers first produced a short descriptive sentence of a radiological image using only the visual features. Then, they attempted to produce more informative descriptions with multiple sentences. However, this introduced new challenges in content selection and ordering [1].

The multi-branch CNN-LSTM fusion network-driven system with the BERT-based semantic evaluator that we describe in this paper has the potential to ensure that relevant findings are included and presented in a clear, ordered format. This solution would help radiologists quickly identify critical information, make more informed decisions, and thus improve the overall efficiency of radiological workflows. Additionally, it could reduce the time required for CT interpretation, allowing for faster diagnosis and treatment of patients, especially emergency room ones. Finally, the computational efficiency of the proposed system makes it particularly suited for those settings where time and resource constraints are paramount.

REFERENCES

[1] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *Artif. Intell. Med.*, vol. 106, Jun. 2020, Art. no. 101878.

[2] D.-R. Beddiar, M. Oussalah, and T. Seppänen, "Automatic captioning for medical imaging (MIC): A rapid review of literature," *Artif. Intell. Rev.*, vol. 56, no. 5, pp. 4019–4076, May 2023.

[3] J. N. Itri and S. H. Patel, "Heuristics and cognitive error in medical imaging," *Amer. J. Roentgenol.*, vol. 210, no. 5, pp. 1097–1105, May 2018.

[4] T. Pang, P. Li, and L. Zhao, "A survey on automatic generation of medical imaging reports based on deep learning," *Biomed. Eng. Online*, vol. 22, no. 1, p. 48, May 2023.

[5] H. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[6] S. A. R. Moezzi, A. Ghaedi, M. Rahmanian, S. Z. Mousavi, and A. Sami, "Application of deep learning in generating structured radiology reports: A transformer-based technique," *J. Digit. Imag.*, vol. 36, no. 1, pp. 80–90, Aug. 2022.

[7] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2497–2506.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Neural Computation*, vol. 9. Cambridge, MA, USA: MIT Press, Nov. 1997, pp. 1735–1780.

[9] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," 2017, *arXiv:1711.08195*.

[10] Y. Xue et al., "Multimodal recurrent model with attention for automated radiology report generation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Granada, Spain. Cham, Switzerland: Springer, Sep. 2018, pp. 457–466.

[11] X. Li, R. Cao, and D. Zhu, "Vispi: Automatic visual perception and interpretation of chest X-rays," 2019, *arXiv:1906.05190*.

[12] M. Sirshar, M. F. K. Paracha, M. U. Akram, N. S. Alghamdi, S. Z. Y. Zaidi, and T. Fatima, "Attention based automated radiology report generation using CNN and LSTM," *PLoS ONE*, vol. 17, no. 1, Jan. 2022, Art. no. e0262209.

[13] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Nov. 2019.

[14] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in COntext (ROCO): A multimodal image dataset," in *Proc. Int. Workshop Large-Scale Annotation Biomed. Data Expert Label Synth.* Cham, Switzerland: Springer, Jan. 2018, pp. 180–189.

[15] A. García Seco de Herrera, C. Eickhof, V. Andrearczyk, and H. Müller, "Overview of the ImageCLEF 2018 caption prediction tasks," in *Proc. Conference Labs Eval. Forum (CLEF)*, vol. 2125, 2018, pp. 1–12.

[16] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[18] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 590–597.

[19] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019.

[20] R. V. Aswiga and A. P. Shanthi, "A multilevel transfer learning technique and LSTM framework for generating medical captions for limited CT and DBT images," *J. Digit. Imag.*, vol. 35, no. 3, pp. 564–580, Jun. 2022.

[21] G.-Y. Kim, B.-D. Oh, C. Kim, and Y.-S. Kim, "Convolutional neural network and language model-based sequential CT image captioning for intracerebral hemorrhage," *Appl. Sci.*, vol. 13, no. 17, p. 9665, Aug. 2023.

[22] M. Ringnér, "What is principal component analysis?" *Nature Biotechnol.*, vol. 26, no. 3, pp. 303–304, Mar. 2008.

[23] S. Mayer, D. Müller, and F. Kramer, "Standardized medical image classification across medical disciplines," 2022, *arXiv:2210.11091*.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," 2016, *arXiv:1603.05027*.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[27] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, pp. 1–7.

[28] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, pp. 26–40, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1674862X19300047

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318.

[30] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

[31] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380.

[32] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," 2016, *arXiv:1612.07600*.

[33] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Jan. 2018, p. 2.

[34] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.

[35] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 196–202.

[36] C. Yin et al., "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 728–737.

[37] R. Ambati and C. R. Dudyala, "A sequence-to-sequence model approach for ImageCLEF 2018 medical domain visual question answering," in *Proc. 15th IEEE India Council Int. Conf. (INDICON)*, Dec. 2018, pp. 1–6.

[38] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in *Machine Learning in Medical Imaging*. Cham, Switzerland: Springer, 2019, pp. 673–680.

[39] G. O. Gajbhiye, A. V. Nandedkar, and I. Faye, "Automatic report generation for chest X-ray images: A multilevel multi-attention approach," in *Computer Vision and Image Processing*. Cham, Switzerland: Springer, 2020, pp. 174–182.

[40] S. Tomassini et al., "On-cloud decision-support system for non-small cell lung cancer histology characterization from thorax computed tomography scans," *Computerized Med. Imag. Graph.*, vol. 110, Dec. 2023, Art. no. 102310.

[41] J. Qiu et al., "Large AI models in health informatics: Applications, challenges, and the future," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 12, pp. 6074–6087, Dec. 2023.