Displays 89 (2025) 103068

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Biomedical text-based detection of colon, lung, and thyroid cancer: A deep learning approach with novel dataset☆

Kubilay Muhammed Sünnetci

Osmaniye Korkut Ata University, Department of Electrical and Electronics Engineering, Osmaniye, Turkey

ARTICLE INFO

Keywords:
Colon cancer
Lung cancer
Thyroid cancer
LSTM
GRU
BiLSTM

ABSTRACT

Pre-trained Language Models (PLMs) are widely used nowadays and increasingly popular. These models can be used to solve Natural Language Processing (NLP) challenges, and their focus on specific topics allows the models to provide answers to directly relevant issues. As a sub-branch of this, Biomedical Text Classification (BTC) is a fundamental task that can be used in various applications and is used to aid clinical decisions. Therefore, this study detects colon, lung, and thyroid cancer from biomedical texts. A dataset including 3070 biomedical texts is generated by artificial intelligence and used in the study. In this dataset, there are 1020 texts labeled colon cancer, while the number of samples labeled lung and thyroid cancer is equal to 1020 and 1030, respectively. In the study, 70 % of the data is used in the training set, while the remaining data is split for validation and test sets. After preprocessing all the data used in the study, word encoding is used to prepare the model inputs. Furthermore, these documents in the dataset are converted into sequences of numeric indices. Afterward, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), LSTM+LSTM, GRU+GRU, BiLSTM+BiLSTM, and LSTM+GRU+BiLSTM architectures are trained with train and validation sets, and these models are tested with the test set. Both validation and test performances of all developed models are determined, and a Graphical User Interface (GUI) software is prepared in which the most successful architecture has been embedded. The results show that LSTM is the most successful model, and the accuracy and specificity values achieved by this model in the validation set are equal to 91.32 % and 95.67 %, respectively. The F1 score value achieved by this model for the validation set is also equal to 91.32 %. The accuracy, specificity, and F1 score values achieved by this model in the test set are equal to 85.87 %, 92.94 %, and 85.90 %, respectively. The sensitivity values achieved by this model for the validation and test set are 91.33 % and 85.88 %, respectively. These developed models both provide comparative results and have shown successful performances. Focusing these models on specific issues can provide more effective results for related problems. Furthermore, the presentation of a user-friendly GUI application developed in the study allows users to use the models effectively.

1. Introduction

It is an indisputable fact that artificial intelligence-based technologies can be developed using numerical, image, audio, video, and text data, and each of these technologies can be utilized in different applications and appeal to various users [1–8]. In particular, nowadays, as a result of the availability of large language models to users [9], knowledge extraction from texts is an important issue and the parts of data understanding, preparation, modelling, and evaluation need to be emphasized [10]. Text mining can be used effectively and efficiently in various tasks such as biomedical, education, agriculture, and business [11]. One of the most notable of these tasks is biomedical text classification, and knowledge discovery from medical data has recently become popular using natural language processing [12]. Herein, it is seen that text-based artificial intelligence models focusing on various biomedical data have been developed in the literature and these, studies using these models are detailed in the following paragraph.

For text mining from biomedical data, in a study discussing the degree of ambiguity, the corpus of sentences for protein interactions is described with appendices. This paper presents a disambiguation process that resolves the ambiguity of terms. In this work, a large resource is presented by bringing together various resources [13]. Given the complexity of NLP, this study provides an overview of the basic concepts of text mining, commonly used techniques, and datasets. Thus, insights

☆ This paper was recommended for publication by Prof Guangtao Zhai.

E-mail address: kubilaysunnetci@osmaniye.edu.tr.

https://doi.org/10.1016/j.displa.2025.103068
Received 10 February 2025; Received in revised form 24 April 2025; Accepted 27 April 2025
Available online 28 April 2025

on how text mining techniques and datasets can be used are presented here [14]. GHS-Net, a hybrid method with a deep learning approach, is proposed to classify various biomedical texts. In this study, the BiLSTM architecture is utilized, and it is noted that it improves the performance metrics of the model [15]. One study describes how machine learning-based NLP models have been applied to databases and analyses 60 studies. The study mentions that these models can improve the quantification of diagnosis and prognosis [16]. In this study, text similarity is emphasized, and models are developed using biomedical data. This study, which aims at multi-task learning, has shown a 2 % improvement in the biomedical field [17]. In a study focusing on pancreatic cancer using Bidirectional Encoder Representations from Transformers (BERT), it is reported that more successful results are obtained compared to traditional binary classification [18]. In this study, which transforms CREGEX into an active learning model, machine learning and transformer algorithms are examined with biomedical-based datasets. In addition, it is stated that 85 % success can be achieved in the study results [19]. A study using machine learning and deep learning-based techniques on three biomedical datasets reports that the cost of manual labelling can be significantly reduced [20]. Focusing on classifying biomedical texts, this study aims at multi-dimensional categorization and states that the results can be practically applied [21]. In a study examining how to adapt the BERT model to biomedical corpora, it is stated that the developed BioBERT outperforms its counterparts [22]. In a study aiming to capture both distributional and relational contexts from texts, the proposed method is tested using various datasets. The results of the study indicate that the proposed system has better performance than its counterparts [23]. Using CREGEX, a biomedical text classifier, this study proposes an approach for binary and multiple classification. Machine learning techniques are also used in this study, and it is mentioned that support vector machines have competitive results [24]. In a study focusing on medical text and disease diagnosis, several models are examined, including convolutional neural networks, transformers, and BERT. In the study results, it is noted that BERT is more successful than the others in terms of performance metrics [25]. In a study aimed at a post-hoc explanation of machine learning techniques commonly used for biomedical text classification, semantic relationships between text and labels are obtained. Herein, explanations showing the decision boundaries of the model are used, and it is stated that it performs successfully in terms of interpretable explanations [26]. For automatic text classification, LSTM is trained using human and veterinary records, and machine learning techniques are also used here. The authors report that they can achieve F1 score values of 91 % and 70 %, respectively [27]. In this study, an architecture called Spark-Text, which includes Apache Spark data streaming and machine learning techniques, is proposed to extract data from articles on cancer. In this study, the proposed model correctly classified a cancer type with about 94 % accuracy [28]. In a study presenting the FREGEX architecture, Smith-Waterman and Needleman-Wunsch sequence alignment algorithms are used. Machine learning techniques are also used in this study, and the performance is reported to be competitive [29]. In this study, a unified neural network is proposed, and convolutional layers are used for feature extraction. In addition, in this study, a bidirectional GRU is used, and the results show that the proposed method is successful in medical texts [30]. BiLSTM technique is used in a study aiming to improve the performance of biomedical text classification. In this study focusing on cardiovascular diseases, successful results are reported according to the literature [31]. A systematic review focusing on text classification and examining 894 articles addresses data-centric and data quality challenges for biomedical text classification [12]. A study using neural networks focuses on ontology-based text classification. The results show that the proposed method is more successful than traditional text classification [32]. In this study employing machine learning algorithms for biomedical text classification, successful classification performance metrics are obtained [33]. In a study using fastText architecture for biomedical sentence classification, an F1 score of 91.7 % has been achieved [34]. In this study, which focuses on text classification for imbalanced data, architectures such as convolutional neural networks, transformers, LSTM, and GRU are used. In this study, which aims to determine the presence and absence of disease, it is reported that successful performances are achieved [35]. Machine learning and deep learning techniques are utilized in this study for clinical text classification. In the results, it is stated that an F1 score of 97 % is obtained for fracture classification [36]. One can see how popular biomedical text classification has been in the past and still is today from a study showing that the vector space model for text classification provides competitive results [37]. In this context, all of these studies focus on biomedical text classification. Its importance and its continuous development can be understood from this paragraph. Studies in the literature that directly focus on colon, lung, and thyroid cancer classification are presented in the following paragraph.

There is a publicly available dataset for colon, lung, and thyroid cancer classification [38]. While this dataset has been referenced in a few studies [39,40], a study focusing on text-based classification for these three cancer types uses various machine learning algorithms. Machine learning techniques are trained after a preprocessing process. In the study results, the accuracy of the proposed method is reported to be equal to about 78 % [41]. In another study training decision trees and random forest algorithms with this dataset, it is noted that 99 % accuracy can be achieved [42]. A study combining support vector machines and BERTopic clustering with explainable inconsistency algorithm utilizes preprocessing techniques. It is stated that 91 % accuracy value can be reached in the study results [43]. Considering these studies, it is significant to note that the proposed method is developed by creating a novel dataset. Although some studies in the literature focus on the classification of biomedical data in various fields [13–37], limited studies focus on colon, lung, and thyroid cancer classification [39–43]. The novel dataset used in this study allows the development of generalizable models. In addition, the fact that this study has software allows it to be used effectively. In addition, various techniques are used in this study. Thus, the comparative performance of the models concerning each other is evaluated. Considering all this information, the proposed method is competitive with the literature, both in terms of performance and the models developed in more specific areas.

In this paper, a novel dataset is created using a specific procedure with the help of ChatGPT. The biomedical texts generated for colon, lung, and thyroid cancer types are employed in the training phase for 70 % of the texts. In contrast, the remaining texts are used for validation and testing. After preprocessing the biomedical texts, LSTM, GRU, BiLSTM, LSTM+LSTM, GRU+GRU, BiLSTM+BiLSTM, and LSTM+GRU+BiLSTM architectures are trained and tested with the novel and preprocessed data. By evaluating the validation and test performances of these seven models, the most successful architecture is determined, and a software application is designed in which this architecture could be used effectively. The models developed in this study contribute significantly to the literature in terms of having competitive results with a novel dataset and successfully detecting colon, lung, and thyroid cancer types from biomedical texts.

2. Data structure

The biomedical texts used in this study are obtained using ChatGPT [44] developed by OpenAI [45]. The free version (GPT-4) is used in this study, and firstly, about 50 questions related to colon, lung, and thyroid cancer are uniquely created by ChatGPT. Herein, some examples of prepared questions are as follows:

Colon Cancer: 1. What measures can be taken to prevent colon cancer? 2. How important is chemotherapy in the treatment of colon cancer? 3. Is there a risk of metastasis in colon cancer? 4. What imaging techniques are used in the diagnosis of colon cancer? 5. Which organs can be affected in the treatment of colon cancer?

Lung Cancer: 1. What are the symptoms of lung cancer? 2. What are

the stages of lung cancer? 3. What is the life expectancy of a patient with lung cancer? 4. How effective is computed tomography scanning in the diagnosis of lung cancer? 5. What are the social rights of lung cancer patients?

Thyroid Cancer: 1. What are the risk factors for thyroid cancer? 2. In which age groups is thyroid cancer more common? 3. How common is thyroid cancer in men? 4. What alternative methods can be used in the treatment of thyroid cancer? 5. How is the quality of life maintained in the treatment of thyroid cancer?

These questions are used to create the dataset for this study and are scrutinized for a novel dataset. Afterward, ChatGPT is usually asked to provide 20 diverse answers for each question. When giving these answers, ChatGPT is often instructed to produce semantically different texts, thus creating biomedical texts. Furthermore, ChatGPT is instructed to write texts with an average word count of 40–100 words in each answer. In this way, a novel dataset is prepared in such a way that both the word lengths are varied, and these texts are different from each other. In the text data labeled colon, lung, and thyroid cancer produced by ChatGPT, the words "colon, lung, and thyroid" are removed from all answers and replaced with a space to avoid overfitting. This is to prevent the model from directly predicting labels from these words and to create a more generalizable model. Considering this information, the dataset information used in the study is presented in the table below.

Table 1 provides the statistics of the biomedical text dataset used in the study. From this table, it can be seen that the numbers of samples used for colon, lung, and thyroid cancers are equal to 1020, 1020, and 1030, respectively. For the architectures developed in the study, the dataset used is split 70 % and 30 % into training and validation & test sets, respectively. In addition, 50 % of this total 30 % of data is used in the validation, while 50 % is used to test the models. In all the models developed, the training, validation, and test sets are the same. Herein, 714 samples labeled colon cancer are used in the training, while 306 samples are used in the validation and test sets. Similarly, these values are equal to 714–306 and 721–309 for samples labeled lung and thyroid cancer, respectively. That is, the training set includes a total of 2149 samples, while the validation and test sets have 461 and 460 samples, respectively.

On the other hand, the word cloud obtained from the texts of the data separated for training is given in Fig. 1. This figure is based on the frequencies of the words in the training text data used in the study. The aim here is to plot the unique elements in the training data with sizes corresponding to their frequencies. In other words, the higher the frequency or frequency of repetition, the larger the word size needs to be in the word cloud. In Fig. 1, it can be seen that the most frequently repeated words are therapy, patients, treatment, and cancer.

Fig. 2 presents a histogram for training document lengths. This figure is a histogram of the document lengths in the training data, and it can be seen that most of the training documents are less than 100 tokens. In this figure, the word counts or token values can be seen, and thus, the dataset used in the study can be better analysed for model training. In this study, the sequences are truncated to have a length of 105 when converting the documents to sequences of numeric indices for fast training and low memory usage.

Table 1
Biomedical text dataset statistics used in the study.

Classes	Training Set (70 %)	Validation (0.5) & Test (0.5) Set (30 %)		Total
Colon Cancer	714	153	153	1020
Lung Cancer	714	153	153	1020
Thyroid Cancer	721	155	154	1030
Total	2149	461	460	3070

3. The proposed BTC Strategy

A biomedical text document is used for colon, lung, and thyroid cancer classification in this study. The novel dataset used in the study is generated by ChatGPT and includes a total of 3070 biomedical texts. This data is split into 70 % training and 30 % validation & test sets. Afterward, preprocessing is applied to all texts in the training, validation, and test sets. Furthermore, word encoding is used to prepare the inputs of the models used in the study, and sequences are truncated by setting a target length.

Therein, with the help of doc2sequence, the documents are converted into sequences of numeric indices. Thus, 2149x1, 461x1, and 460x1 cell arrays are obtained for the training, validation, and test sets, respectively. Training and validation sets are used to train seven different models used in the study. These architectures are layers1, layers2, layers3, layers4, layers5, layers6, and layers7. They include the LSTM, GRU, BiLSTM, LSTM+LSTM, GRU+GRU, BiLSTM+BiLSTM, and LSTM+GRU+BiLSTM layers. Fully connected, softmax, and classification layers are used in these architectures. After the training, the trained models are saved, and performance metrics are calculated. In addition, all the proposed architectures are tested using a test set, and the performance of these models is also analysed. A user-friendly GUI application with the most successful trained model embedded in it is designed for this study. Based on this information, a block diagram showing all the steps of the method proposed in this study is presented in Fig. 3.

3.1. Preparing dataset

This section describes in detail how the novel dataset used in the study has been prepared for training. First, the dataset is split into training, validation, and test sets. In this study, the ratio of the training set is 70 %, while the total distribution of validation and test sets is equal to 30 %. Furthermore, the ratio of validation and test sets is set to 50 % and 50 %, respectively. The training, validation, and test sets used in the study are split randomly. All three datasets include descriptions, and the documents of these datasets are tokenized using tokenizedDocument. Afterward, the contents of the tokenized documents are converted to lowercase with the lower function. Finally, punctuations are erased from these documents using the erase punctuation function. This process is completed for the dataset used in the study. Therefore, preprocessing steps are applied separately to the training, validation, and test sets. This process provides the appropriate data for the models' inputs. In light of this information, the preprocessing block used in the study is presented in Fig. 4. After the preparation of the dataset, the architectures used are explained in the following section.

3.2. Data classification Approaches

This section presents the classifiers used in this study. Seven different classifiers, layers1, layers2, layers3, layers4, layers5, layers6, and layers7 are used in the study. layers1 has a LSTM [46,47] architecture, while layers2 has a GRU [48,49]. On the other hand, layers3 has the BiLSTM [47,50,51] structure, while layers4 includes two consecutive LSTMs. layer5 and layer6 have two consecutive GRU and BiLSTM layers, respectively. Lastly, layer7 has LSTM, GRU, and BiLSTM architectures in a sequential manner. The layers of the classification architectures used in this study are given in Table 2. The table indicates that the first component of all architectures is the sequence input layer, followed by the word embedding layer. The differences of these classifiers occur in the third, fourth, and fifth layers. After that, fully connected and softmax layers are available for all these architectures. The value used for a fully connected layer equals three in this study. Finally, these architectures are created using the classification layer. The connection structures of these architectures can be seen in Table 2.

The LSTM, GRU, and BiLSTM architectures and their diverse variants are used for biomedical text document classification. Thus, various

Fig. 1. Word cloud of the training text data.

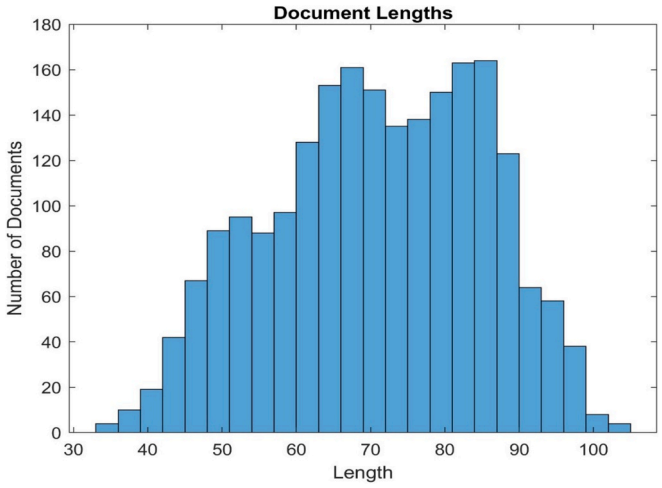

Fig. 2. Histogram for the training document lengths.

architectures are trained, and their differences and performance metrics are evaluated in the study. The following section presents a user-friendly GUI application developed to make it easier for users to employ and where the most successful model is embedded.

3.3. User Interface Design

The title of this GUI developed within the scope of this study is MedicalTextPredictionApp. This application is designed in a very simple way so that users can use it effectively. The designed user-friendly GUI application has a text panel. By using this panel, users can write the relevant texts here with the help of the keyboard. In this panel, when the current line block is full, text can be written to the following line as standard. All of this information provided by the user can be examined with the help of a scroll bar. There is also a classification panel in this GUI application. This panel reports the predictions of the most successful model proposed in the study. In this study, since the best performance metrics are obtained with the LSTM architecture, it is embedded in this user-friendly GUI application. In order for this model to predict the text label, users need to click on the click to predict the text label button. Thus, the text label can be provided in the developed GUI application. Additionally, users can close this GUI application by clicking the close button. In this GUI application, when the click to predict the text label button is clicked, the model takes about 1 s to predict and print on the screen. With this GUI application, users can quickly and effectively predict biomedical text. Based on this information, screenshots of this user-friendly GUI application developed in this study are presented in Fig. 5.

3.4. Model evaluation metrics

This section describes the classification metrics used to analyze the performance of the proposed models. The classification metrics calculated for the architectures layers1, layers2, layers3, layers4, layers5, layers6, and layers7 (LSTM, GRU, BiLSTM, LSTM+LSTM, GRU+GRU, BiLSTM+BiLSTM, and LSTM+GRU+BiLSTM) are accuracy, sensitivity, specificity, precision, F1 score, and MCC values. Accuracy is obtained by dividing the number of correct predictions by the total number of samples [52]. This metric is necessary but not sufficient for the evaluation of classifier performance. Herein, it is reasonable to calculate and analyse the sensitivity and specificity metrics. The sensitivity metric, recall, is calculated as the number of true positive predictions divided by the total number of positives and ranges from 0 to 1 [53]. In specificity, the number of true negative predictions is divided by the total number of negatives and ranges from 0 to 1 [52,53]. The precision metric is the number of true positive predictions divided by the total number of positive predictions [53,54]. The F1 score is obtained by calculating the harmonic mean of precision and sensitivity [52,55]. Matthews Correlation Coefficient (MCC) represents the correlation between observed and predicted classifications and is sensitive to imbalanced data and ranges from −1 to 1 [54]. In the light of this information, the formulas of these metrics used in this study can be written as follows [52–55],

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

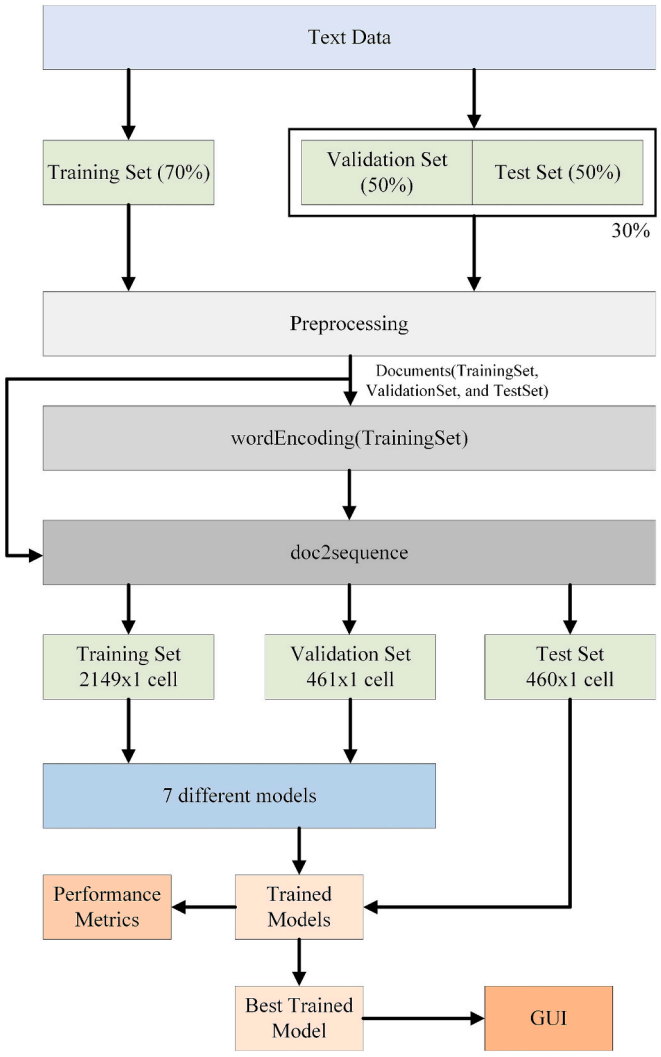

Fig. 3. Flowchart of the proposed method.

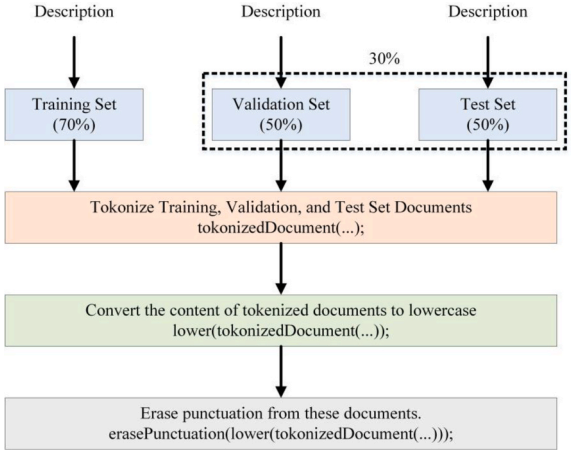

Fig. 4. Preprocessing block used in the proposed method.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{F1 } Score = \frac{2.Precision.Sensitivity}{Precision + Sensitivity} \tag{5}$$

$$\text{MCC} = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

4. Experimental Configuration

This section describes the confusion matrices obtained for LSTM, GRU, BiLSTM, LSTM+LSTM, GRU+GRU, BiLSTM+BiLSTM, and LSTM+GRU+BiLSTM architectures, the training parameters of the models, and the training progress of the most successful model. The computer used in this study has Intel(R) Core(TM) i5-6400 CPU @ 2.70 GHz 8.00 GB RAM (x64).

Fig. 6 indicates the validation and test confusion matrices obtained for layers1, layers2, layers3, layers4, layers5, layers6, and layers7 models in the study. When the validation and test confusion matrices of the layer1 architecture are examined, it is seen that it successfully predicts 144-139-138 and 138-129-128 biomedical texts for the texts labeled colon, lung, and thyroid cancer, respectively. For layers2, these values are 142-113-144 and 142-110-139, while in layers3, the number of correctly predicted biomedical texts are 132-136-130 and 122-135-127. In the layers4 architecture, 34 texts labeled thyroid cancer are predicted incorrectly for the validation set, while this value is 37 in the test set. This model successfully predicts the labels of 140–145 and 130–141 texts from the validation and test sets for texts labeled colon and lung cancer. On the other hand, layers5, layers6, and layers7 architectures successfully predict 131-109-147&127-113-139, 136-129-136&130-125-133, and 136-114-137&132-108-132 texts from the colon, lung, and thyroid labeled texts in the validation and test sets, respectively.

Table 3 presents the training parameters for the layers1, layers2, layers3, layers4, layers5, layers6, and layers7 architectures used in this study. Note that the same parameters are used for all the models developed in this table. The Optimizer is set to “adam” for all models used in the study. In training these models, values of 80 and 150 are used for Embedding Dimension and Hidden Units, respectively. Moreover, the Mini Batch Size, Gradient Threshold, Validation Frequency, Gradient Decay Factor, and Squared Gradient Decay Factor values for these models equal 8, 2, 200, 0.9, and 0.9990, respectively. On the other hand, the Epsilon and Initial Learn Rate values for the models are 10^{-8} and 10^{-4}, respectively. Learn Rate Drop Factor, Learn Rate Drop Period, and L2Regularization values equal 0.1, 10, and 10^{-4}, respectively. For all models, the Gradient Threshold Method and Objective Metric Name are set to l2norm and loss, respectively. Batch Normalization Statistics, Output Network, and Acceleration are set to auto.

Fig. 7 shows the training progress for layer1, the most successful model in the study. In these figures, training accuracy, validation accuracy, training loss, and validation loss are plotted according to iterations. It can be seen from the accuracy-based figure that the performance of the model generally increases in both training and validation. The loss values obtained in the training of this model decrease meaningfully with iteration. These figures show that the models have successful performances. The validation accuracy and loss values obtained after 8040 iterations are equal to 91.3232104 and 0.4676875, respectively. In this figure, the time required to train the model is 12 min and 52 s. Multiplying 30 epochs and 268 ’iterations per epoch’ yields 8040 maximum

Table 2
Layers of the architectures used for classification in the study.

	layers1	layer2	layer3	layers4	layers5	layers6	layers7
1.	Sequence Input Layer						
2.	Word Embedding Layer						
3.	lstmLayer	gruLayer	bilstmLayer	lstmLayer	gruLayer	bilstmLayer	lstmLayer
4.				lstmLayer	gruLayer	bilstmLayer	gruLayer
5.							bilstmLayer
6.	Fully Connected Layer						
7.	Softmax Layer						
8.	Classification Layer						

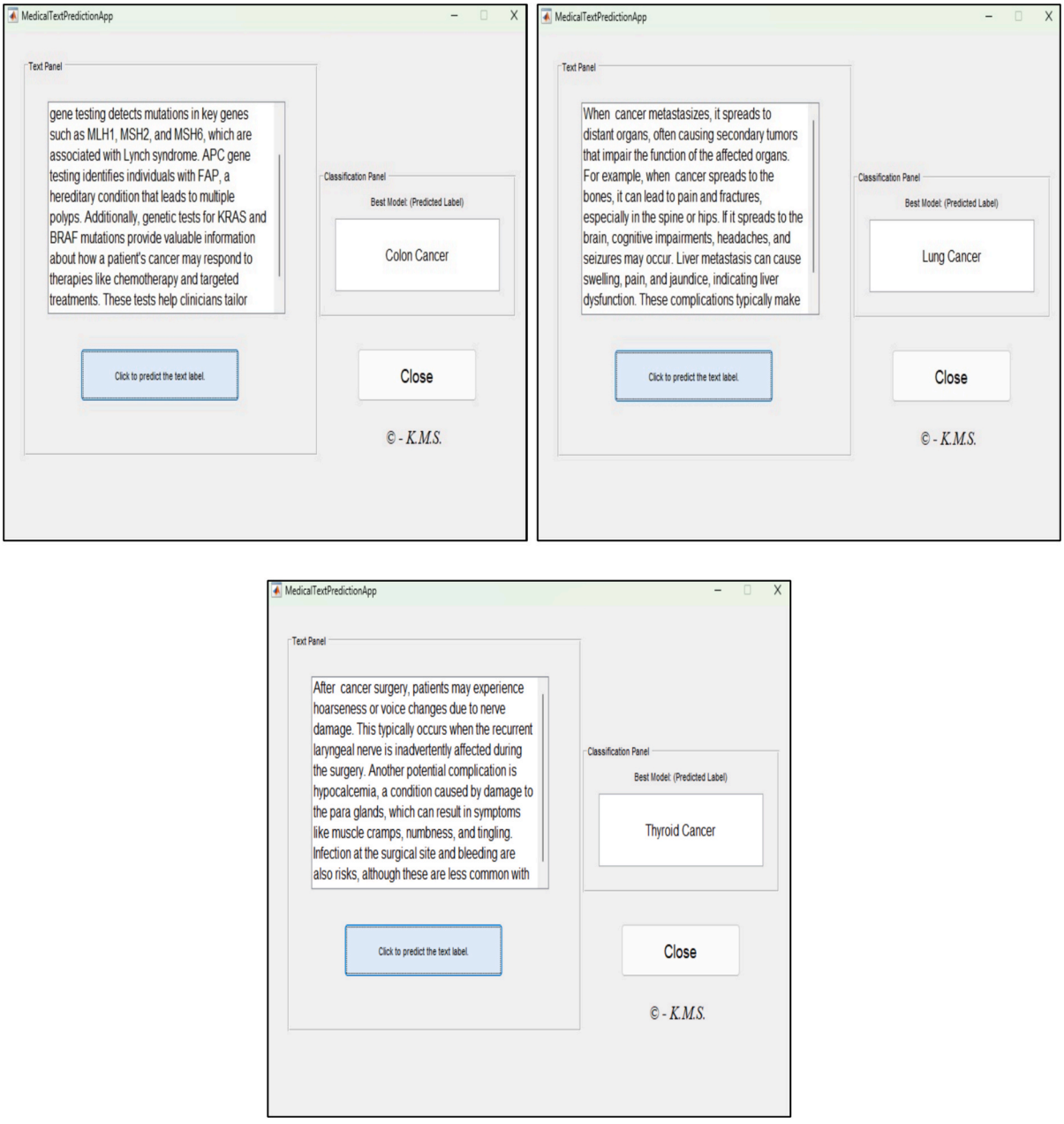

Fig. 5. Screenshots of disease predictions for different texts of the GUI application designed in the study.

iterations. In this figure, the learning rate is also shown, and it is seen that hardware resource and learning rate schedule are Single CPU and constant, respectively.

5. Results and Discussion

In this part, the results of the LSTM, GRU, BiLSTM, LSTM+LSTM, GRU+GRU, BiLSTM+BiLSTM, and LSTM+GRU+BiLSTM architectures developed using a novel dataset are presented and the success of the

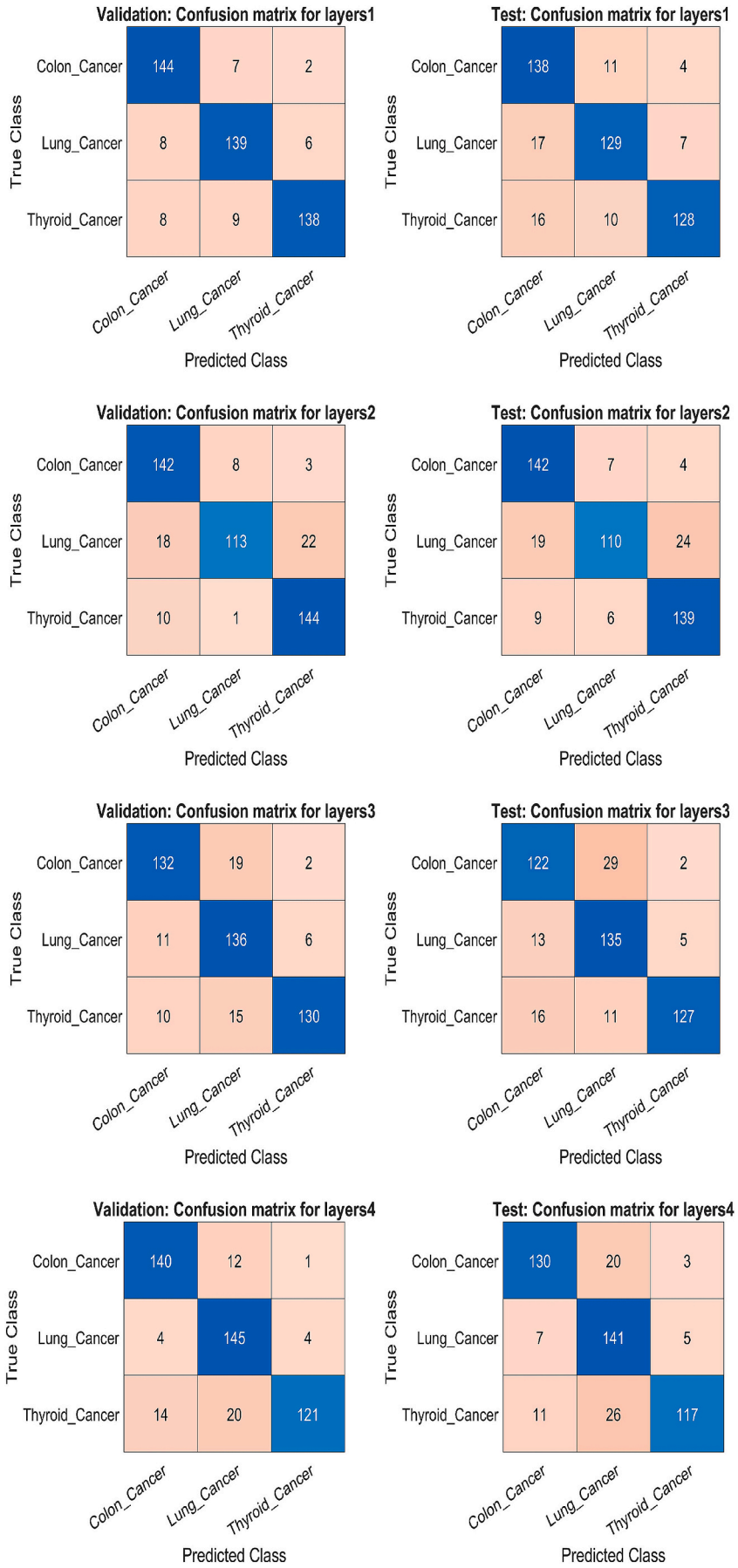

Fig. 6. Validation and test set confusion matrices obtained for the models trained in the study.

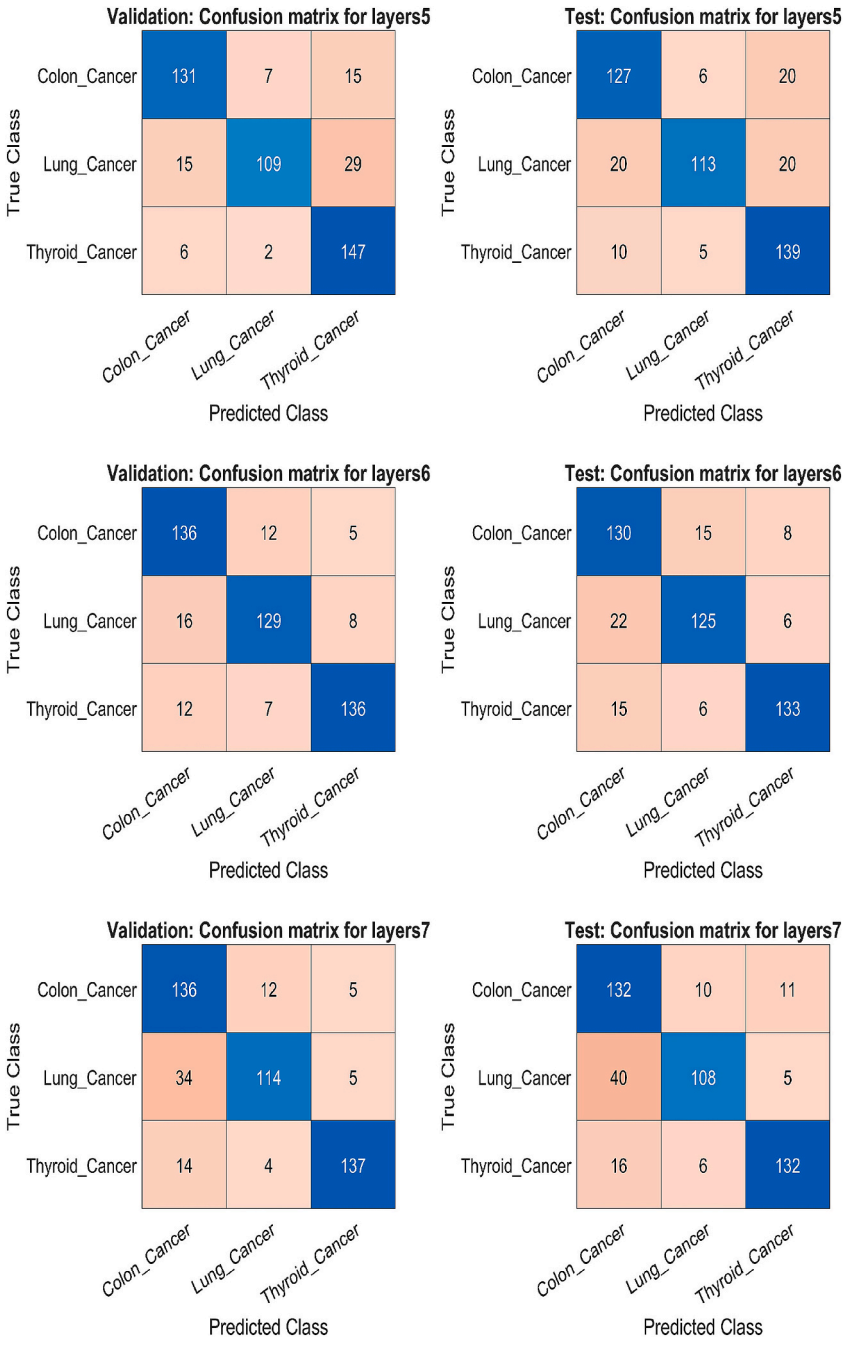

Fig. 6. (*continued*).

study compared to the literature is examined according to these results. Herein, the classification performance metrics achieved for validation and test sets from the models developed in the study are presented, and these metrics are obtained separately from the validation and test confusion matrices.

Table 4 presents the performance metrics for the seven different architectures developed in the study. First, the accuracy values achieved by layers1, i.e. the LSTM architecture, for the validation and test sets are 91.32 % and 85.87 %, respectively. Similarly, the precision and F1 score values that this model can achieve in validation and test sets are 91.40 %-91.32 % and 86.26 %-85.90 %, respectively. In addition, this model can reach 87.03 % MCC value in the validation set. The accuracy, precision, and F1 score values that the GRU architecture named layers2 can achieve for the test set are 85 %, 85.40 %, and 84.75 %, respectively. These values that the BiLSTM-based layers3 architecture can achieve for the test set are 83.48 %, 84.24 %, and 83.59 %, respectively. In this context, it is seen that LSTM is the best among the single-structure architectures. On the other hand, for layers4 containing two consecutive LSTMs, these values are equal to 84.35 %, 85.61 %, and 84.40 %, while for layers5 containing two consecutive GRUs, they are equal to 82.39 %, 83.22 %, and 82.34 %. Additionally, in layers6 containing two consecutive BiLSTM structures, the accuracy, precision, and F1 score values obtained from the test set are 84.35 %, 84.65 %, and 84.41 %, respectively. Finally, in layers7, which contains LSTM, GRU, and BiLSTM structures sequentially, these values obtained from the test set are 80.87 %, 82.17 %, and 80.94 %. Herein, it is seen that less success is achieved by using LSTM, GRU, and BiLSTM in a hybrid way compared to others. When all models are evaluated together, it is seen that the most successful model is layers1, namely LSTM. When the results are examined, it is seen that this model can be trained both quickly and cost-effectively

Table 3
Training parameters for the models used in the study.

Parameter Selection	*layers1, layers2, layers3, layers4, layers5, layers6, and layers7*
Mini Batch Size	8
Optimizer	adam
Embedding Dimension	80
Hidden Units	150
Initial Learn Rate	10^{-4}
Validation Frequency	200
Gradient Decay Factor	0.9
Epsilon	10^{-8}
Learn Rate Drop Factor	0.1
Learn Rate Drop Period	10
L2Regularization	10^{-4}
Gradient Threshold Method	l2norm
Objective Metric Name	loss
Batch Normalization Statistics	auto
Output Network	auto
Acceleration	auto

in terms of computational resources.

Considering these results, the proposed method can be compared with the literature. First, this study differs from the literature in that it uses a novel dataset. Herein, the creation of the dataset with the help of ChatGPT according to a specific protocol and the removal of colon, lung, and thyroid from the biomedical texts to avoid overfitting allows the models to perform the learning process on a challenging problem. In addition to the use of artificial intelligence-based models in various fields [56–65], detecting cancer types from biomedical texts is a significant problem. In this context, the introduction presents studies in the literature focusing on biomedical text mining [13–37]. On the other hand, there are studies using a publicly available dataset to detect colon, lung, and thyroid cancer from biomedical texts. However, there are 996 unique values in this dataset. Therefore, it is thought that this dataset needs to be improved and edited. However, it can be seen that the performance of the models developed in [41–43] can reach 99 %. Independently of this study, the BiLSTM technique has achieved 99.12 % accuracy, 99.23 % sensitivity, 99.56 % specificity, 99.14 % precision, and 99.17 % F1 score on the publicly available dataset. In the proposed method, an accuracy of 91.32 % is achieved with a novel dataset. With the help of the GUI application developed in this study, which contains competitive results compared to the literature, users can use the developed models effectively and efficiently. It is seen that the methods proposed in this study can provide successful results even in difficult problems such as text mining. Hereabouts, it is noted that the performance metrics obtained from the methods trained using a novel dataset have successful and applicable results compared to the literature. Considering this information, this study stands out because it uses a novel dataset, focuses directly on colon, lung, and thyroid cancers, and has successful performance metrics. The models developed in this study can also be provided for expert use to classify the specified cancer types, saving labor and cost. PLMs have been used actively and effectively,

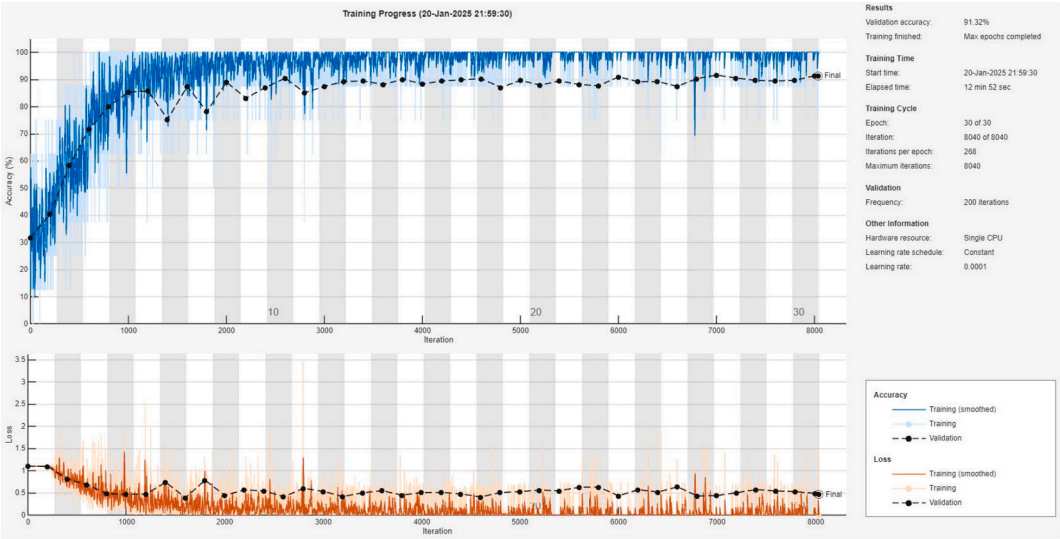

Fig. 7. Training progress for the best model in the study.

Table 4
Performance metrics achieved for seven different models.

Models	Metrics	Accuracy	Sensitivity	Specificity	Precision	F1 score	MCC
layers1	Val.	0.9132	0.9133	0.9567	0.9140	0.9132	0.8703
	Test	0.8587	0.8588	0.9294	0.8626	0.8590	0.7901
layers2	Val.	0.8655	0.8652	0.9327	0.8712	0.8633	0.8017
	Test	0.8500	0.8499	0.9250	0.8540	0.8475	0.7777
layers3	Val.	0.8633	0.8634	0.9318	0.8683	0.8641	0.7975
	Test	0.8348	0.8348	0.9175	0.8424	0.8359	0.7560
layers4	Val.	0.8807	0.8811	0.9404	0.8885	0.8801	0.8257
	Test	0.8435	0.8437	0.9218	0.8561	0.8440	0.7721
layers5	Val.	0.8395	0.8390	0.9196	0.8517	0.8377	0.7660
	Test	0.8239	0.8237	0.9119	0.8322	0.8234	0.7403
layers6	Val.	0.8698	0.8698	0.9350	0.8712	0.8700	0.8055
	Test	0.8435	0.8434	0.9218	0.8465	0.8441	0.7666
layers7	Val.	0.8395	0.8393	0.9198	0.8493	0.8400	0.7644
	Test	0.8087	0.8086	0.9044	0.8217	0.8094	0.7198

especially in recent times. These models can support users in a variety of areas, but by focusing them on specific areas, it is possible to achieve meaningful results in more specific areas. Therefore, this study aimed to classify the determined cancer types. In addition, in this study, architectures such as LSTM, GRU, and BiLSTM are used individually and sequentially. LSTM, GRU, BiLSTM, LSTM+LSTM, GRU+GRU, BiLSTM+BiLSTM architectures as well as LSTM+GRU+BiLSTM architecture are used to evaluate various features of LSTM, GRU, and BiLSTM together. When all these architectures are analyzed, the fact that the proposed method has comparative results gives an idea about how the models can perform for this dataset.

6. Concluding Remarks

This paper successfully detects colon, lung, and thyroid cancer from biomedical text documents. The study utilizes a novel dataset of 3070 biomedical texts labeled as colon, lung, and thyroid cancer. While 70 % of the data is used for training, 30 % is used for validation and test sets. Of the 30 % of the data, 50 % is split for the validation set and 50 % for testing. Herein, cell-based training, validation, and test sets are obtained with the help of preprocessing, word encoding, and doc2seqeunce. LSTM, GRU, BiLSTM, LSTM+LSTM, GRU+GRU, BiLSTM+BiLSTM, and LSTM+GRU+BiLSTM models are trained and validated using the obtained training and validation sets. Afterward, all models are tested using the test set, and both validation and test performance metrics are calculated for these models. The most successful model is identified in the calculation results, and a software is created using this model. It is seen that the models developed in the study provide satisfactory results according to the literature, and the performance of the models is functional for biomedical text document classification. On the other hand, it can be noted that this study, which focused on the detection of colon, lung, and thyroid cancer, can also focus on different cancer types. Moreover, the document lengths of the developed models used for training can be further increased in future studies, and more advanced models can be obtained by using different architectures.

CRediT authorship contribution statement

Kubilay Muhammed Sünnetci: Investigation, Methodology, Software, Writing – original draft, Conceptualization, Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank ChatGPT for its support in preparing the data.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] J. Luo, D. Zhu, D. Li, Classification-enhanced LSTM model for predicting river water levels, J. Hydrol. 650 (2025) 132535.
[2] K.M. Sunnetci, A. Alkan, Biphasic majority voting-based comparative COVID-19 diagnosis using chest X-ray images, Expert Syst. Appl. 216 (2023) 119430.
[3] U. Bal, A. Bal, Ö.T. Moral, F. Düzgün, N. Gürbüz, A deep learning feature extraction-based hybrid approach for detecting pediatric pneumonia in chest X-ray images, *Phys. Eng. Sci. Med.* 47 (1) (2024) 109–117.
[4] K.M. Sunnetci, F.E. Oguz, M.N. Ekersular, N.G. Gulenc, M. Ozturk, A. Alkan, Comparative bladder cancer tissues prediction using vision transformer, *J. Imag. Inform. Med.* (2024).
[5] J.W. Chang, H.S. Ma, Z.Y. Hu, Multi-Level Transfer Learning using Incremental Granularities for environmental sound classification and detection, *Appl. Soft Comput.* 169 (2025) 112619.
[6] Y. Zahid, C. Zarges, B. Tiddeman, J. Han, Adversarial diffusion for few-shot scene adaptive video anomaly detection, *Neurocomputing* 614 (2025) 128796.
[7] A. Riyadi, M. Kovacs, U. Serdült, V. Kryssanov, Benchmarking with a language model initial selection for text classification tasks, *Mach. Learn. Knowl. Extr.*. 2025 (2025) 1–25.
[8] E. Akkuş, U. Bal, F.Ö. Koçoğlu, S. Beyhan, Hyperparameter optimization of pre-trained convolutional neural networks using adolescent identity search algorithm, *Neural Comput. & Applic.* 36 (2024) 1523–1537.
[9] F.E. Oguz, M.N. Ekersular, K.M. Sunnetci, A. Alkan, Can chat GPT be utilized in scientific and undergraduate studies? *Ann. Biomed. Eng.* 52 (5) (2024) 1128–1130.
[10] A. Hotho, A. Nürnberger, G. Paaß, A brief survey of text mining, *J. Language Technol. Comput. Linguistic* 20 (1) (2005) 19–62.
[11] V. Veeramachaneni, Large language models: a comprehensive survey on architectures, applications, and challenges, *Adv. Innov. Comp. Program. Languages* 7 (1) (2025) 20–39.
[12] C.Y.Y. Kesiku, A. Chaves-Villota, B. Garcia-Zapirain, Natural language processing techniques for text classification of biomedical documents: a systematic review, *Information (Switzerland)* 13 (10) (2022) 499.
[13] F. Rinaldi, K. Kaljurand, R. Sætre, Terminological resources for text mining over biomedical scientific literature, *Artif. Intell. Med.* 52 (2) (2011) 107–114.
[14] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, B. Shen, Biomedical text mining and its applications in cancer research, *J. Biomed. Inform.* 46 (2) (2013) 200–211.
[15] M.A. Ibrahim, M.U. Ghani Khan, F. Mehmood, M.N. Asim, W. Mahmood, GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification, *J. Biomed. Inform.* 116 (2021) 103699.
[16] N. Mollaei, C. Cepeda, J. Rodrigues, H. Gamboa, Biomedical Text Mining: Applicability of Machine Learning-based Natural Language Processing in Medical Database, in: *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies*, Biostec, 2022, pp. 159–166.
[17] Y. Peng, Q. Chen, Z. Lu, An empirical study of multi-task learning on BERT for biomedical text mining, *arXiv* (2020) 1–10, arXiv:2005(02799v1).
[18] J. He, L. Rasmy, D. Zhi, C. Tao, Advancing pancreatic cancer prediction with a next visit token prediction head on top of Med-BERT, *arXiv* (2025) 1–17, arXiv:2501 (02044).
[19] C.A. Flores, R.L. Figueroa, J.E. Pezoa, Active learning for biomedical text classification based on automatically generated regular expressions, *IEEE Access* 9 (2021) 38767–38777.
[20] U. Naseem, M. Khushi, S.K. Khan, K. Shaukat, M.A. Moni, A comparative analysis of active learning for biomedical text mining, *Appl. Syst. Innov.* 4 (1) (2021) 23.
[21] H. Shatkay, F. Pan, A. Rzhetsky, W.J. Wilbur, Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users, *Bioinformatics* 24 (18) (2008) 2086–2093.
[22] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (4) (2020) 1234–1240.
[23] M.A. Parwez, M. Fazil, M. Arif, M.T. Nafis, M.R. Auwul, Biomedical text classification using augmented word representation based on distributional and relational contexts, *Comput. Intell. Neurosci.* 2023 (1) (2023) 2989791.
[24] C.A. Flores, R.L. Figueroa, J.E. Pezoa, Q. Zeng-Treitler, CREGEX: A biomedical text classifier based on automatically generated regular expressions, *IEEE Access* 8 (2020) 29270–29280.
[25] Z. Chen, Enhancing medical text classification and disease diagnosis with BERT advances in deep learning for healthcare, *J. Comp. Sci. Software Appl.* 5 (1) (2025) 1–9.
[26] M. Moradi, M. Samwald, Explaining black-box models for biomedical text classification, *IEEE J. Biomed. Health Inform.* 25 (8) (2021) 3112–3120.
[27] G.R. Venkataraman, A.L. Pineda, O.J. Bear Don't Walk, A.M. Zehnder, S. Ayyar, R. L. Page, C.D. Bustamante, M.A. Rivas, FasTag: Automatic text classification of unstructured medical narratives, *PLoS One* 15 (6) (2020) 1–18.
[28] Z. Ye, A.P. Tafti, K.Y. He, K. Wang, M.M. He, SparkText: biomedical text mining on big data framework, *PLoS One* 11 (9) (2016) 1–15.
[29] C.A. Flores, R.L. Figueroa, J.E. Pezoa, FREGEX: A Feature Extraction Method for Biomedical Text Classification using Regular Expressions, in: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 6085–6088.
[30] L. Qing, W. Linhong, D. Xuehai, A novel neural network-based method for medical text classification, *Future Internet* 11 (12) (2019) 255.
[31] N. Ahmed, F. Dilmaç, A. Alpkocak, Classification of biomedical texts for cardiovascular diseases with deep neural network using a weighted feature representation method, *Healthcare (Switzerland)* 8 (4) (2020) 392.
[32] S. Malik, S. Jain, Knowledge-infused text classification for the biomedical domain, *Int. J. Inform. Syst. Model. Design* 13 (10) (2022) 1–15.
[33] B. Behera, G. Kumaravelan, Performance evaluation of machine learning algorithms in biomedical document classification, *Int. J. Adv. Sci. Technol.* 29 (5) (2020) 5704–5716.
[34] A. Agibetov, K. Blagec, H. Xu, M. Samwald, Fast and scalable neural embedding models for biomedical sentence classification, *BMC Bioinf.* 19 (2018) 541.

[35] H. Lu, L. Ehwerhemuepha, C. Rakovski, A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance, *BMC Med. Res. Method.* 22 (2022) 181.
[36] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E.J. Atkinson, S. Amin, H. Liu, A clinical text classification paradigm using weak supervision and deep representation, *BMC Med. Inf. Decis. Making* 19 (2019) 1.
[37] M. Lan, C.L. Tan, J. Su, H.B. Low, Text representations for text categorization: a case study in biomedical domain, *Int. Joint Conf. Neural Networks* 2007 (2007) 2557–2562.
[38] Medical Text Dataset-Cancer Doc Classification @ www.kaggle.com. https://www.kaggle.com/datasets/falgunipatel19/biomedical-text-publication-classification.
[39] A.T.M. Rony, M. Shariful Islam, T. Sultan, S. Alshathri, W. El-Shafai, MediGPT: exploring potentials of conventional and large language models on medical data, *IEEE Access* 12 (2024) 103473–103487.
[40] A. Bojesomo, M. Seghier, L. Hadjileontiadis, A. AlShehhi, Revolutionizing disease diagnosis with large language models : a systematic review revolutionizing disease diagnosis with large language models: a systematic review, *Research Square* 1 (2024) 1–32.
[41] E. Kucuk, I. Cicek, Z. Kucukakcali, C. Yetis, Comparative analysis of machine learning algorithms for biomedical text document classification: A case study on cancer-related publications, *Med. Sci.* 13 (1) (2024) 171–174.
[42] H. Oktavianto, H.W. Sulistyo, G. Wijaya, D. Irawan, G. Abdurrahman, Analisis komparasi kinerja metode decision tree dan random forest dalam klasifikasi teks data kesehatan, *Bina Insani ICT Journal* 11 (1) (2024) 56–65.
[43] P. Mortezaagha, A. Rahgozar, Inconsistency Detection in Cancer Data Classification Using Explainable-AI, medRxiv (2024) 1–18.
[44] ChatGPT @ https://chatgpt.com/. https://chat.openai.com/.
[45] OpenAI @ https://openai.com/. https://openai.com/index/chatgpt/.
[46] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, *Neural Comput.* 31 (7) (2019) 1235–1270.
[47] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780.
[48] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv* (2014) 1–9, arXiv:1412(3555v1).
[49] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, *arXiv* (2014) 1–15, arXiv:1406(1078v3).
[50] S. Khan, M. Fazil, V.K. Sejwal, M.A. Alshara, R.M. Alotaibi, A. Kamal, A.R. Baig, BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection, *J. King Saud University – Comp. Inform. Sci.* 34 (7) (2022) 4335–4344.
[51] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Netw.* 18 (5–6) (2005) 602–610.
[52] M. Hossin, M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, *Int. J. Data Mining Knowl. Manage. Process* 5 (2) (2015) 1–11.
[53] Ž. Vujović, Classification model evaluation metrics, *Int. J. Adv. Comput. Sci. Appl.* 12 (6) (2021) 599–606.
[54] A. Tharwat, Classification assessment methods, *Appl. Comput. Inf.* 17 (1) (2021) 168–192.
[55] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification an overview, *Arxiv* (2020) 1–17, arXiv:2008(05756v1).
[56] K.M. Sunnetci, E. Kaba, F.B. Celiker, A. Alkan, Deep network-based comprehensive parotid gland tumor detection, *Acad. Radiol.* 31 (1) (2024) 157–167.
[57] K.M. Sunnetci, S. Ulukaya, A. Alkan, Periodontal bone loss detection based on hybrid deep learning and machine learning models with a user-friendly application, *Biomed. Signal Process. Control* 77 (2022) 103844.
[58] K.M. Sunnetci, Ö. Aydin, A. Alkan, Determination of methylene violet concentration using classification algorithms, *Iran J. Comp. Sci.* (2024) 1–12.
[59] F.E. Oğuz, A. Alkan, Weighted ensemble deep learning approach for classification of gastrointestinal diseases in colonoscopy images aided by explainable AI, Displays 85 (2024) 102874.
[60] M.N. Ekersular, A. Alkan, Detection of COVID-19 disease with machine learning algorithms from CT images, *Gazi Univ. J. Sci.* 37 (1) (2024) 169–181.
[61] M. Balcı, A. Alkan, Identification of wart treatment evaluation by using optimum ensemble based classification techniques, *Biomed. Signal Process. Control* 95 (Part A) (2024) 106491.
[62] K.M. Hasib, M. Oli Ullah, M. Imran Nazir, A. Akter, M. Saifur Rahman, ICDP: An Improved Convolutional Neural Network Model to Detect Pneumonia from Chest X-Ray Images, Springer Nature Singapore, Singapore, 2024, pp. 467–479.
[63] N. Islam, K.M. Hasib, M.F. Mridha, S. Alfarhood, M. Safran, M.K. Bhuyan, Fusing global context with multiscale context for enhanced breast cancer classification, *Sci. Rep.* 14 (1) (2024) 27358.
[64] M.A. Rahman, M.I. Masum, K.M. Hasib, M.F. Mridha, S. Alfarhood, M. Safran, D. Che, GliomaCNN: an effective lightweight CNN model in assessment of classifying brain tumor from magnetic resonance images using explainable AI, *CMES - Computer Modeling in Engineering and Sciences* 140 (3) (2024) 2425–2448.
[65] M.S.H. Shovon, M.F. Mridha, K.M. Hasib, S. Alfarhood, M. Safran, D. Che, Addressing uncertainty in imbalanced histopathology image classification of HER2 breast cancer: an interpretable ensemble approach with threshold filtered single instance evaluation (SIE), *IEEE Access* 11 (2023) 122238–122251.