Heliyon 10 (2024) e32279

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Deep-GenMut: Automated genetic mutation classification in oncology: A deep learning comparative study

Emad A. Elsamahy [a], Asmaa E. Ahmed [a,*], Tahseen Shoala [b], Fahima A. Maghraby [a]

[a] College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Cairo, Egypt
[b] Environmental Biotechnology Department, College of Biotechnology, Misr University for Science and Technology, Giza, 12563, Egypt

ARTICLE INFO

Keywords:
BERT
BioBERT
LSTM
BiLSTM
Genetic mutation
Text classification
Cancer detection
Deep learning

ABSTRACT

Early cancer detection and treatment depend on the discovery of specific genes that cause cancer. The classification of genetic mutations was initially done manually. However, this process relies on pathologists and can be a time-consuming task. Therefore, to improve the precision of clinical interpretation, researchers have developed computational algorithms that leverage next-generation sequencing technologies for automated mutation analysis. This paper utilized four deep learning classification models with training collections of biomedical texts. These models comprise bidirectional encoder representations from transformers for Biomedical text mining (BioBERT), a specialized language model implemented for biological contexts. Impressive results in multiple tasks, including text classification, language inference, and question answering, can be obtained by simply adding an extra layer to the BioBERT model. Moreover, bidirectional encoder representations from transformers (BERT), long short-term memory (LSTM), and bidirectional LSTM (BiLSTM) have been leveraged to produce very good results in categorizing genetic mutations based on textual evidence. The dataset used in the work was created by Memorial Sloan Kettering Cancer Center (MSKCC), which contains several mutations. Furthermore, this dataset poses a major classification challenge in the Kaggle research prediction competitions. In carrying out the work, three challenges were identified: enormous text length, biased representation of the data, and repeated data instances. Based on the commonly used evaluation metrics, the experimental results show that the BioBERT model outperforms other models with an F1 score of 0.87 and 0.850 MCC, which can be considered as improved performance compared to similar results in the literature that have an F1 score of 0.70 achieved with the BERT model.

1. Introduction

A gene mutation can be defined as a change in the normal sequence of deoxyribonucleic acid (DNA) that makes up a gene. This change is different from the DNA sequence that is common in most people [1,2]. Gene mutations can vary in size and have a significant impact on multiple DNA components, potentially involving a large stretch of chromosomes that includes multiple genes [3]. Tumors are typically heterogeneous, and their genomic profiles typically contain different types of genetic mutations [4–6]. These mutations can be caused by errors in DNA replication during cell division or by environmental influences and radiation [7,8]. Consequently, these

* Corresponding author.
E-mail addresses: e.elsamahy@aast.edu (E.A. Elsamahy), A.Morsy13121@student.aast.edu (A.E. Ahmed), Tahsen.shoala@must.edu.eg (T. Shoala), fahima@aast.edu (F.A. Maghraby).

https://doi.org/10.1016/j.heliyon.2024.e32279
Received 12 January 2024; Received in revised form 30 May 2024; Accepted 31 May 2024
Available online 31 May 2024

defective patterns can have a significant impact in the development of serious diseases such as cancer [9–12]. The identification and analysis of cancer-causing genes is crucial in the field of clinical trials [13]. Currently, physicians have to manually evaluate and categorize individual mutated genes by analyzing the information contained in the clinical literature in text format [14,15]. However, although the manual process of interpreting genomics for gene classification is critical to saving lives [16,17], it is challenging as it is both time-consuming and subjective. Furthermore, successful cancer treatment depends largely on the discovery of mutated genes [18]. To address this problem, Memorial Sloan Kettering Cancer Center (MSKCC) has created a repository of precise oncology information based on multiple mutations carefully captured by leading scientists and physicians [19]. Some researchers achieved good results by analyzing textual evidence using Natural Language Processing (NLP) approaches integrated with pre-trained models and deep learning (DL) methods [20,21]. Consequently, this study attempts to use DL-based models to categorize gene mutations based on textual evidence, which will increase the effectiveness and speed of cancer tumor diagnosis compared to manual techniques. Although traditional statistical machine learning (ML) models may outperform manual approaches in text categorization, they still require human efforts for feature extraction, resulting in increased labor costs and challenges in obtaining effective features [22,23]. In this context, previous literature results show that DL methods outperform traditional statistical machine-learning techniques. Moreover, using these methods circumvents the challenges of labor-intensive manual feature engineering and potentially reduces the costs associated with implementation [24]. The emergence of novel technologies in the medical field has facilitated the accumulation of a significant amount of data on cancer, thereby promoting advances in medical research. Medical researchers have pointed out that predicting cancer outcomes is one of the most difficult and important challenges. Various efforts have been made in this research area based on the MSKCC dataset. In 2018, Gangmin et al. used Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction with two different classification models: XGBoost and support vector machine (SVM). The results presented in the paper showed that XGBoost outperformed SVM. In addition, the use of confusion metrics proved that XGBoost have advanced prediction ability [25]. In 2019, Jingwen Xu et al. utilized various neural network-based structures for classifying gene mutations. In this work, a parallel hybrid CNN-BiGRU neural network was presented, which demonstrated superior performance compared to standard classifiers based on neural network models [26]. In 2020, Samruddhi Mhatre et al. exploited various machine learning methods, including K-Nearest Neighbors, Naive Bayes, Random Forest (RF) Classifier, Logistic Regression (LR), Linear Support Vector Machine, and Stacking Model, to categorize genetic variations using an expert-annotated knowledge base. TF-IDF, a hot encoding and label encoding technique, was used for feature preprocessing. The KNN algorithm achieved the lowest log loss value, however logistic regression (with adjustment) was chosen as the best model because KNN produced an overfitted value [27].

In 2020, Yuhan Su et al. exploited BERT to categorize genetic mutations by analyzing textual evidence from an annotated database. The BERT model's ability to classify complex clinical text has been improved. Among the three BERT-based models, the BERT plus abstract shortness method was found to be superior as a standalone prototype, achieving an F1 score of 0.705 [8]. In 2021 Meenu Gupta et al. recommended various classifiers to categorize genetic mutations based on medical evidence. CountVectorizer, TF-IDF Vectorizer, and Word2Vec were utilized as text transformers. The efficiency of the proposed approach was evaluated using different ML classification models, namely LR, RF, an Extreme Gradient Boosting (XGB) classifier, and a Recurrent Neural Network (RNN). Empirical results showed that the RNN using the Word2Vec text transformation model achieved the highest accuracy that reached 71 % [28]. In 2023, Sanad Aburass and colleagues created a hybrid ensemble model that combines LSTM, BiLSTM, CNN, GRU, and GloVe embeddings. This model was intended to handle the task of classifying gene mutations in cancer using Kaggle's Personalized Medicine dataset and it achieved an F1-score of 0.831 [29].

Despite the notable progress in genetic mutation classification, a substantial gap exists in the current literature. The complexity and heterogeneity of tumors, coupled with the diverse nature of genetic alterations, pose significant challenges in accurately categorizing gene mutations. While previous studies have made commendable efforts to utilize various machine learning and deep learning techniques, there remains a need for further improvement in accuracy. The existing approaches, including TF-IDF, XGBoost, CNN-BiGRU, and BERT-based models, have showcased varying degrees of success. However, the quest for enhanced accuracy and efficiency persists. Therefore, this study aims to bridge the existing gap by employing BioBERT and other deep learning models to elevate the precision and speed of cancer tumor diagnosis. The goal is to surpass the limitations of labor-intensive manual processes and achieve a more robust and reliable classification of genetic mutations.

The proposed work presented significant contributions to the field of gene mutation classification, addressing key challenges, and advancing the state-of-the-art techniques in several aspects. Firstly, the utilization of BioBERT represents a novel approach to classifying genetic mutations. By applying BioBERT to the categorization of gene mutations, this study explores the potential benefits of leveraging specialized language models in the biomedical domain. Secondly, the work focuses on enhancing the accuracy of classification models for gene mutations. Through the utilization of deep learning-based approaches, specifically BioBERT, the precision and reliability of automated mutation analysis were improved. This contributes to the ongoing efforts to develop more effective and efficient methods for gene mutation categorization, addressing the limitations associated with manual interpretation and classification. Thirdly, the comparative analysis between BERT, BioBERT as well as LSTM and BiLSTM provides insights into the capabilities of BioBERT specifically in the context of biomedical data classification. By highlighting the advantages of BioBERT over BERT, the study contributes valuable information for researchers and practitioners seeking optimal models for genetic mutation analysis. This comparison addresses a gap in the literature, providing a nuanced understanding of the performance variations between these widely used language models. The subsequent sections of this article are structured in the following manner: Section 2 presents the contextual background about the used algorithms within this study. Section 3 illustrates the materials and methods used in the study as well as the results achieved. The article's conclusion is depicted in Section 4.

2. Background

This section provides a comprehensive overview of the necessary background information on the used techniques. Let's start by exploring transformers [30], which are an integral part of BERT and BioBERT [39]. Next, an examination of BERT and BioBERT is discussed as they serve as a fundamental part of our methodology. Next, word embedding techniques such as word2vec, one hot encoding, TF-IDF and Glove are introduced, which are necessary steps before using the LSTM and BiLSTM classifiers. It is worth noting that BERT and BioBERT contain embedded word embedding techniques.

2.1. *Transformers*

The transformer [30] is a Neural Network (NN) construction originally designed for translation tasks, as shown in Fig. 1. There are 6 blocks present in both the encoder and the decoder, each of which comprises double coatings. The early coating is a multi-skull care coating, though the additional is a completely linked coating. To ensure a smooth flow of information, residual connections are employed, connecting the involvement of a coating to its production. Additionally, a coating norm is applied afterward respectively to normalize the data. The attention layer assigns three vectors to each word: a value (v), a query (q), and a key (k). The extent of dependency concerning an agreed word (w_s) and a goal word (w_t) is measured via taking the dot product of the word's query (q_{ws}) and the goal word's key (k_{wt}). Thus, the word is signified by captivating a biased regular of the values of the goal words using (q_{ws}, k_{wt}) as the respective weights. For normalization, the influences are additionally accustomed through a softmax function to ensure that they sum up to one before the weighted average is computed. To simplify the notation, the solutions of the goal words are organized in matrix (K), enquiries in matrix (Q), and the values of the target words in concatenated matrix (V). The final output of the layer is then determined using Equation (1).

$$\text{Attention}\,(\text{Q},\text{K},\text{V}) = softmax\left(\frac{(Qk^T)}{\sqrt{dk}}\right)\text{V} \tag{1}$$

where (dk) represents the dimensionality of the vector (k).

In the computation of attention-based word representation presented above, one can observe that the positional information of words in the target sequence is lost due to the weighted average across targeted words. To address this issue, a solution is to augment

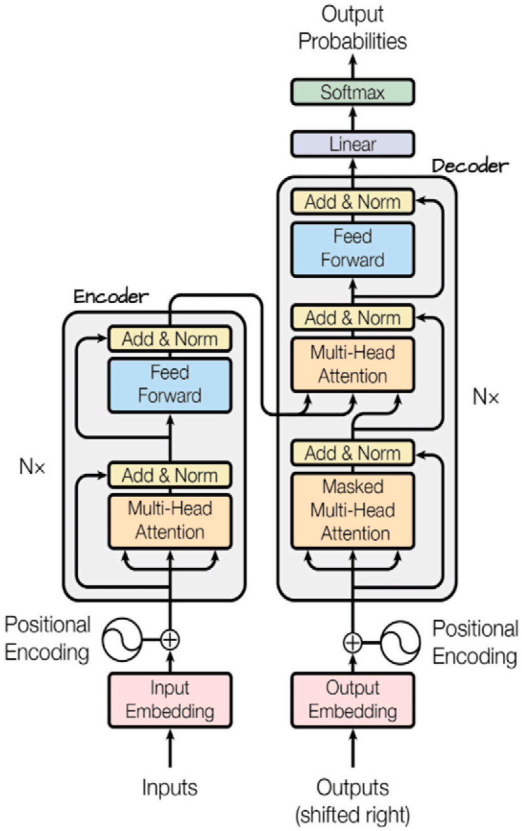


Fig. 1. Transformer architecture [30].

each word representation with a vector that encodes its position (n), known as positional encoding (PE). The PE vector is derived by means of Equation (2).

$$PE_{(pos,k)} = \left\{ \sin\ \sin\left(\frac{pos}{10^{\frac{4k}{d}}k}\right), if\ k\ is\ even, \cos\ \cos\left(\frac{pos}{10^{\frac{4k}{d}}k}\right), if\ k\ is\ odd, \right. \tag{2}$$

where (k) is the variable representing the guide of the calculated rate for the positional vector, whereas (pos) signifies the location of the expression.

2.2. BERT/BioBERT

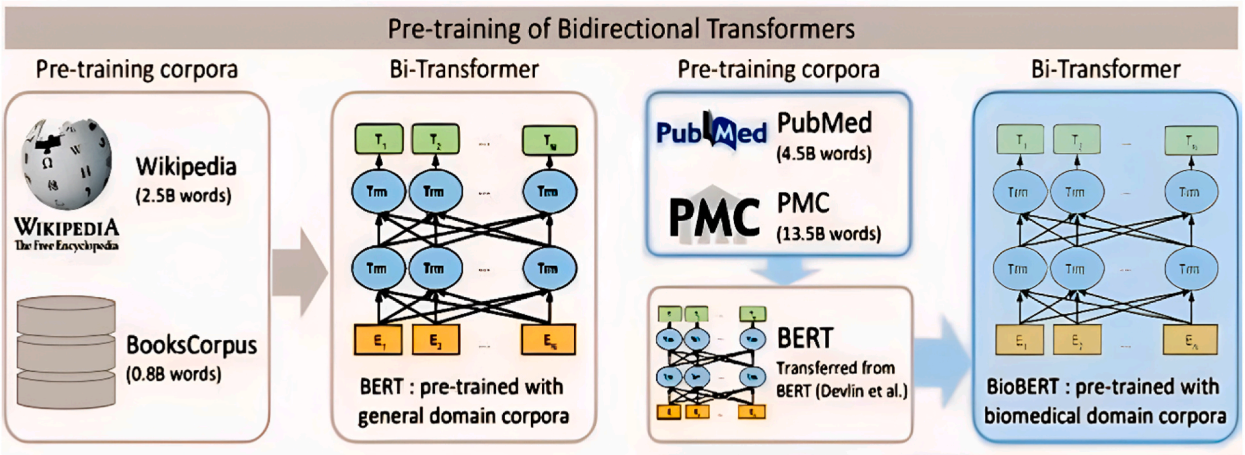
BERT is an acronym for Bidirectional Encoder Representations from Transformers. Its purpose is to perform pre-training on unlabeled text and create rich bidirectional understanding by considering both previous and subsequent contexts. Consequently, adding just one more output layer to the BERT model enables the creation of state-of-the-art models for a wide range of natural language processing tasks. On the other hand, BERT's pre-training includes a significant amount of unlabeled text, which includes the entire Wikipedia and book corpus words. BioBERT is a specialized language representation model developed for the field of biomedical text mining [31]. Thanks to the pre-training of large biomedical corpora, it outperforms the BERT model in several tasks related to the analysis of biomedical texts. Pre-training BERT on such corpora improves its ability to capture complex biomedical content, as shown in Fig. 2(a). By leveraging BioBERT's contextual embeddings and semantic representations, researchers and practitioners can extract meaningful insights from biomedical text data as shown in Fig. 2(b), accelerate biomedical research, and ultimately contribute to advancements in healthcare and life sciences.

2.3. LSTM/BiLSTM

The Long Short-Term Memory (LSTM) model has gained wide acceptance in the NLP field due to its exceptional ability to acquire knowledge from sequential inputs [32]. The LSTM architecture effectively minimizes the challenge of gradient vanishing and exploding that is common with traditional recurrent neural networks (RNNs). This is achieved by using memory blocks with gates instead of hidden vectors, allowing long-term memory to be maintained. The mechanism used in LSTM networks consists of three different layers, namely the input gate, the forget gate, and the output gate as shown in Fig. 3(a). These gates are regulated by the sigmoid function. The memory cells within the LSTM architecture can update and display their stored information only when deemed essential. LSTM has been shown to be effective in various areas to achieve state-of-the-art results. BiLSTM is an extension of LSTM that processes sequential input in both directions, as shown in Fig. 3(b).

2.4. GloVe word embedding

Extracting features from text is a crucial step in machine learning, especially for unstructured data such as text datasets. In 2014, J. Pennington developed the concept of word embedding by creating Global Vectors for Word Representation (GloVe) [34], a vector representation of learning spaces for words that was integrated into Stanford's NLP laboratory [32,33]. This technique involves converting words into vector representations within a contextual framework to solve the challenge of capturing contextual word associations in a computationally tractable feature space. This technique is utilized in our work for word embedding as it introduced



(a)BERT

(b)BioBERT

Fig. 2. Structure and pre-training procedure for (a) BERT and (b) BioBERT [31].

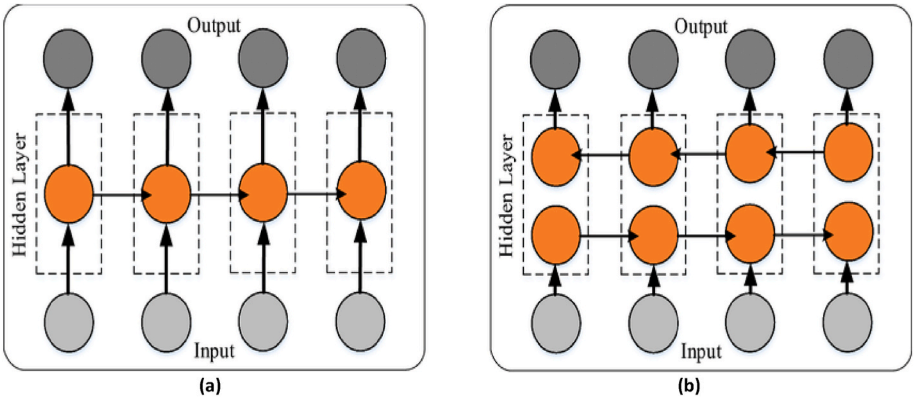


Fig. 3. Architectures of (a) LSTM and (b) BiLSTM [42].

superior results when used in conjunction with LSTM and BiLSTM in classification tasks [32].

3. Materials and methods

In this section, four models are utilized to classify genetic mutations. These models are BERT and BioBERT for feature extraction and classification (referred to as Models 1 and 2). In addition, since LSTM and BiLSTM have no feature extraction phase, they are utilized in conjunction with Glove for feature extraction (referred to as Models 3 and 4). The dataset obtained from the Kaggle competition [15] based on the (MSKCC) dataset was exploited as the input for the four aforementioned models for training and testing purposes.

3.1. Description of dataset

The MSKCC dataset, comprising 3320 records, is divided into two distinct files: one for variants data, while the other for text data. The variants data file contains details about gene mutations, including various fields such as a unique identifier (ID) linking variations to clinical textual evidence, the gene harboring the DNA change (Gene), the alteration in amino acids resulting from the mutation (Variation), and the genetic mutation's category (Class). On the other hand, the text data file encompasses clinical descriptions utilized in the classification of genetic mutations, categorized into nine different classes. Sample entries from the dataset are presented in Table 1.

While analyzing the dataset, the following aspects were considered:

1. There are differences in the lengths of text between classes. multiple classes contain shorter terms, while others include redundant descriptions. To overcome this issue, three truncation methods are employed to truncate clinical text (head – head + tail - middle) that extract crucial information using various techniques. In head truncation, the initial 512 token of the text is retained, while the latter part is discarded. This strategy prioritizes information at the beginning of the text. While head + tail truncation aims to preserve both the beginning and end of the text by removing a portion from the middle. This strategy balances information from both ends of the text. For instance, using the same sentence, head + tail truncation may retain the initial 256 tokens and the final 256 tokens. Middle truncation is a technique where the text is split into two approximately equal halves, and then the first 512 tokens from the second part are retained.
2. This approach helps us avoid any adverse effects caused by the substantial disparities in text lengths as well as mitigating any negative impacts arising from significant variations in the lengths of textual content.
3. The dataset exhibits imbalanced sample sizes across different classes, as shown in Fig. 4. To address this discrepancy, long records were split into smaller chunks where certain classes had fewer records compared to others. Consequently, these chunks were processed individually, and the outputs were then combined with max pooling. This approach aimed to mitigate the imbalances

Table 1
Samples of the dataset (after merging the two files).

ID	Text	Gene	Mutation (variation)	Class
1	Abstract Retinoblastoma is a pediatric retinal tumor… ….	TMEM216	G77A	4
2	Autosomal dominant deficiency of signal transducer and activator of transcription 3 (STAT3) is the main… ….	STAT3	F384L	2
3	here is a paucity of information about the molecular perturbations involved in MPM……… ….	EGFR	W731L	7
4	NRF2 is a transcription factor that mediates stress responses. Oncogenic mutations in NRF2…… ….	KEAP1	Deletion	1

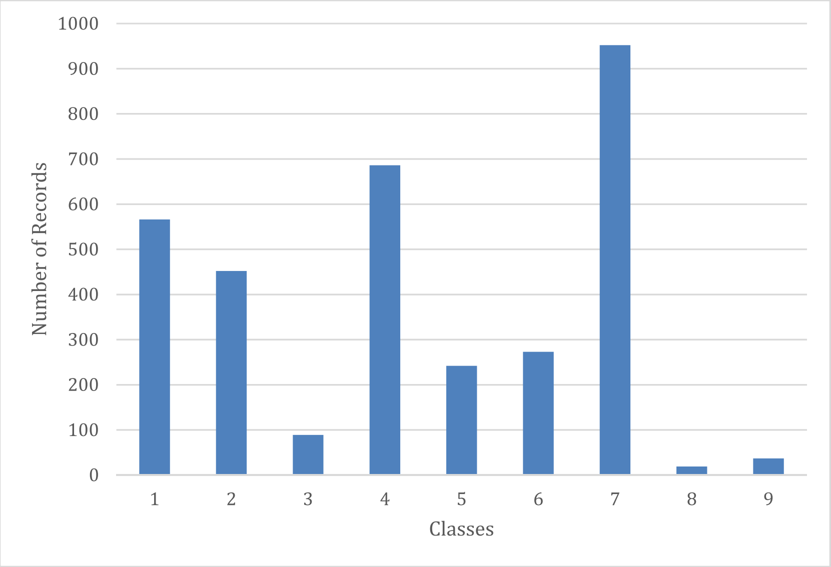


Fig. 4. Number of records in the nine classes before balancing the dataset.

between categories as it showed good results as mentioned in the literature [36]. Fig. 5 shows the distribution of the nine classes after balancing.

After applying the balancing technique, the number of records was increased to 5176 compared to the original number which is 3320. Next, the dataset was divided into three partitions: training (60 %), validation (20 %), and testing (20 %) which is used to evaluate the effectiveness of the proposed models. The training and validation subsets contained 4170 records, while the testing subset included 1006 records.

3.2. Methodology

In this section, the proposed method leverages advanced natural language processing techniques for gene mutation classification, employing both the BERT/BioBERT models as well as the Glove-based LSTM/BiLSTM models as shown in Fig. 7.

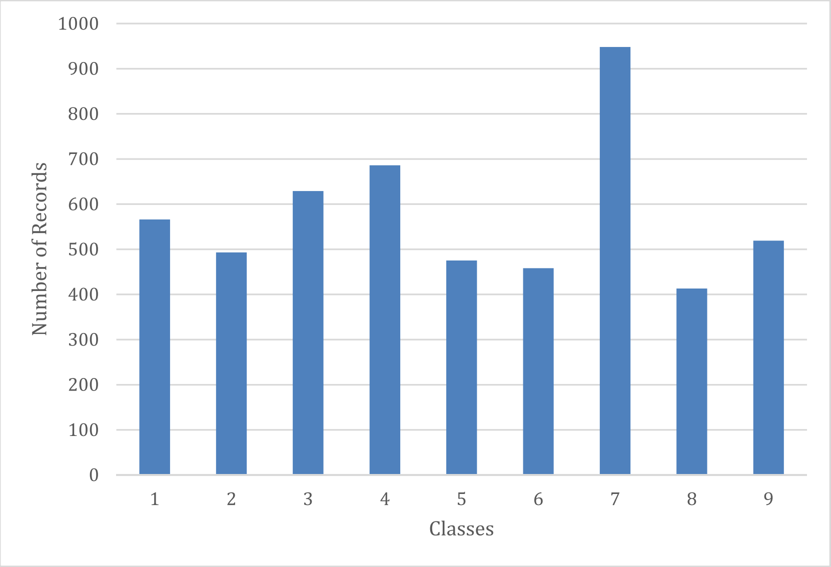


Fig. 5. Number of records in the nine classes after balancing the dataset.

3.2.1. BERT/BioBERT model

Both the BERT/BioBERT models, which share the same architecture but is trained on different data, were constructed using python programming language. In 2023 Alexander Turchin et al. proposed a study aimed to assess the precision of identifying various biomedical concepts in medical narratives by contrasting different versions of the BERT model [40]. The results showed that both BERT/BioBERT proved themselves in classification tasks that are based on biomedical data with high performance. Since, these two models have a maximum input length of 512, therefore, the truncation method was employed to match the length limitation (see section 3.1.1). The BERT/BioBERT models can efficiently transform each word in the text into a one-dimensional vector by retrieving the corresponding word vector from a repository and feeding it into the model as mentioned in Fig. 6. The model's input comprises three distinct sections: token embeddings, segmentation embeddings, and position embeddings. The Adam optimizer [37] is utilized to adaptively update the momentum and learning rate for better performance. while adjusting the batch size to 32, the learning rate to 0.00002, and experimenting with 50 epochs. The batch size and number of epochs were selected based on the literature. The learning rate was deduced based on an ablation study (see section 3.5), as being a crucial hyperparameter that significantly impacts the model's performance. Due to BERT's advanced pretraining and strong generalization capabilities, it is possible to link the output layer of BERT with the external layer to successfully perform downstream tasks. In this case, the output layer is connected to the softmax function for task classification.

3.2.2. LSTM/BiLSTM model

In 2022 Chandra Bhushana Rao Kill et al. employed a technique involving utilizing a neural network known as LSTM and word embedding features using Global Vector (GloVe). The study employed different text datasets to classify fake news. The outcomes highlighted the viability of this system [41]. Based on the success achieved by LSTM and BiLSTM, Glove-based LSTM and BiLSTM were utilized for gene mutation classification, knowing that LSTM processes the data in one direction (from left to right). In contrast, BiLSTM processes the data in both directions (from left to right and from right to left) [38]. That's why BiLSTM exhibits better performance than the LSTMAs shown in Fig. 7, the methodology is divided into sequential stages that are started with preprocessing the dataset. As aforementioned, the data is balanced in nature. Consequently, this distribution indicates that the model may exhibit bias towards certain categories within the dataset. Therefore, up-sampling is used to balance data and remove stop words. Subsequently, the work came to tokenization which involves the segmentation of a given text into individual phrases that are then divided into typographic tokens. Following that feature extraction plays a vital role in machine learning, particularly when dealing with text data. Text datasets are inherently unstructured, requiring techniques that can bring meaning and structure to be utilized by machine learning algorithms. addressing the challenge of capturing contextual relationships between words in a computable feature space. For this study, embedding from the GloVe model, characterized by a dimensionality of 600 was used as input features for models based on Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) architectures based on literature review. The models use a synthesis of hyperparameters, including the embedding matrix derived from preprocessing the GloVe feature, Rectified Linear Unit (Relu) and Hyperbolic Tangent (Tanh) activation functions for hidden gates, softmax activation for output gates, Adam optimizer, a dropout rate of 0.3 and a learning rate of 0.001 based on ablation study (see section 3.5), which controls the speed at which the model learns. The learning rate determines the frequency of weight updates during training.

3.3. Evaluation metrics

The evaluation metrics used to evaluate our proposed classification models are described in this section. The testing phase for the proposed models was carried out using the testing subset. The goal of each model is to correctly identify the class of mutation present in the given text. Four evaluation criteria are described here: accuracy, recall, precision, and F1-score [35]. These metrics are calculated as described in Equations (3)–(7):

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{3}$$

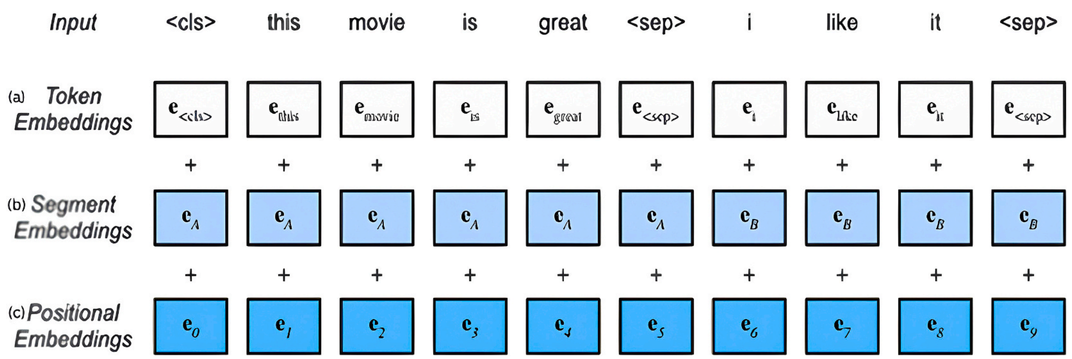


Fig. 6. BERT input representation [39], (a) Token Embeddings, (b) Segment Embeddings, (c) Positional Embeddings.

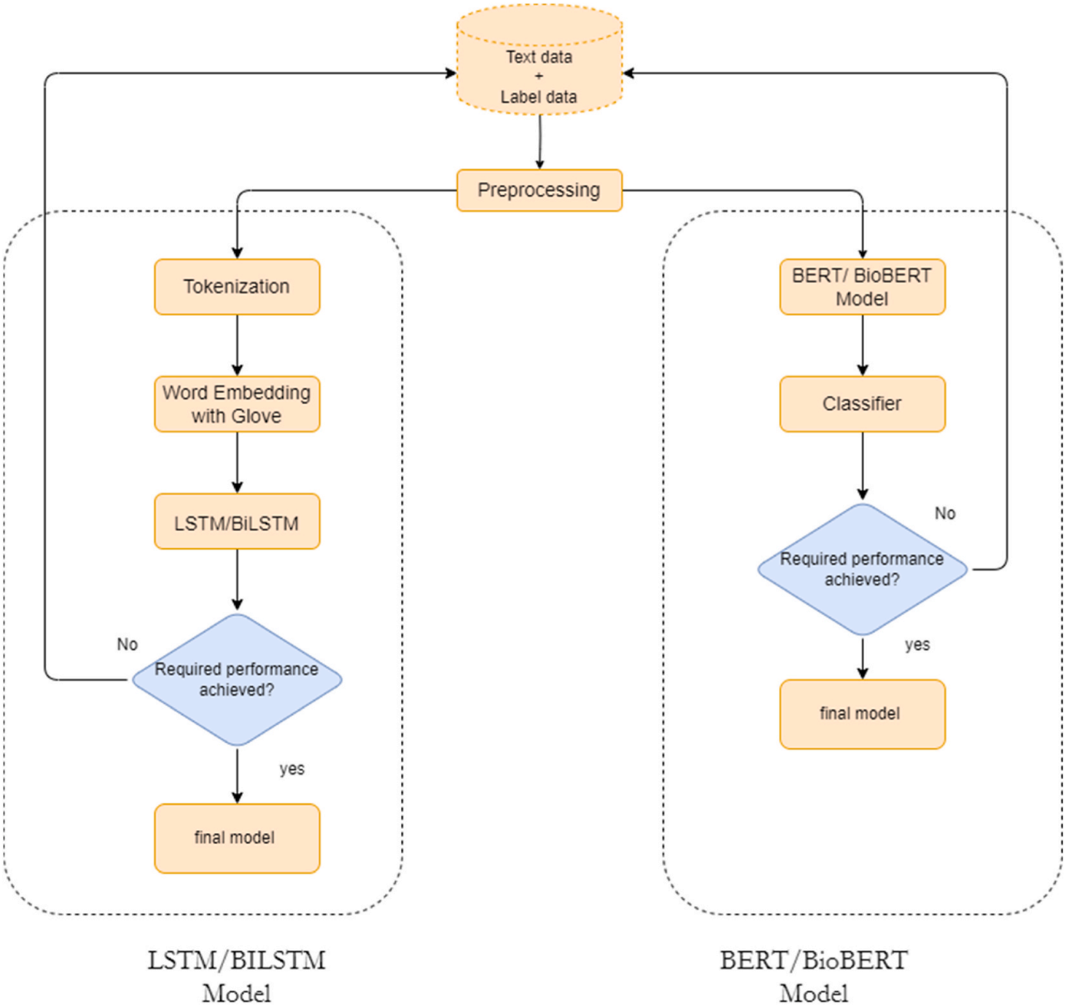


Fig. 7. The proposed framework.

$$Precesion = \frac{T_P}{T_P + F_P} \tag{4}$$

$$Recall = \frac{T_P}{T_p + F_N} \tag{5}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

Table 2
Results achieved using BERT, BioBERT, LSTM, and BiLSTM.

Model	Truncation Method	Performance			
		F1-score	Accuracy	Log loss	MCC
BioBERT	Head	0.783	0.780	0.574	0.731
	Middle	0.721	0.717	0.965	0.695
	Head + Tail	**0.87**	**0.864**	**0.489**	**0.850**
BERT	Head	0.705	0.701	0.642	0.652
	Middle	0.710	0.703	0.966	0.663
	Head + Tail	0.706	0.702	0.882	0.601
LSTM	-	0.643	0.643	0.673	0.592
BiLSTM	-	0.651	0.652	0.672	0.578

$$\mathrm{MCC} = \frac{T_P \,.\, T_N \;-\; F_P \,.\, F_N}{\sqrt{(T_P + F_P)\,.\;(T_P + F_N)\,.\;(T_N + F_P)\,.\,(T_N + F_N)}} \tag{7}$$

where (T_P) is the true positive instances, (T_N) is the true negative instances, (F_P) is the false positive instances, and (F_N) is the false negative instances.

3.4. Results and discussion

In this section, the achieved results using the proposed models are presented. Table 2 shows the results produced by the four models. It is worth noting that, as the dataset exhibits some imbalances, the F1-score will faithfully evaluate the models' performances rather than only using accuracy which may contain some bias. Based on the shown results, the BioBERT model using the three truncation methods has achieved the highest accuracy and F1 score compared to the results exhibited by the other three models. Moreover, using the Head + Tail truncation method showed the best results with the BioBERT model.

This model achieves 0.87 F1-score, 0.850 MCC, and 0.86 accuracy in spite of the limitation of the extreme shortage of training data.

Fig. 8 shows the confusion matrix for the BioBERT model's prediction using the Head + Tail truncation method against the testing data subset. The results illustrated that the model succeeded in classifying the gene mutations for classes 3, 8, and 9 with accuracy over 90 %, classes 1, 4, 6, and 7 with accuracy over 86 %, while classes 2 and 5 were classified with accuracy over 75 %.

In Table 3, a comprehensive comparison between our proposed models against other models presented in the literature for the purpose of genetic mutation classification.

It can be seen that the BioBERT model with the Head + Tail truncation method outperforms all other models either that proposed in this paper as well as others found in the literature based on the accuracy and F1-score against the testing data subset.

3.5. Hardware configuration

All the proposed work in the paper was implemented on Google Colab Pro, a cloud-based platform offering enhanced resources for machine learning and data analysis tasks. The hardware configuration utilized in the work is summarized as follows:

Runtime Environment: Google Colab Pro provided access to high-RAM runtime environments, allowing for efficient handling of memory-intensive tasks. This was particularly beneficial for processing and analyzing our models.

GPU Acceleration: The research leveraged the GPU support offered by Google Colab Pro (A100 GPU). The availability of more powerful GPU resources significantly expedited the training of deep learning models, enhancing the overall efficiency of experimentation.

3.6. The ablation study

In the context of classifying genetic mutations, an ablation study helps to understand the importance of different components or parameters in obtaining accurate results. Based on the information provided, it involves conducting multiple experiments with different parameters. The parameter studied is the learning rate, which turns out to be one of the variables that most influence the model's accuracy. Table 4 (for BioBERT and BERT) shows the resulting accuracies of the proposed models over three learning rate values that have been most commonly used in the literature. The Adam optimizer with 50 epochs was utilized in all experiments. It is worth noting that the Adam optimizer and the 50 training epochs were chosen to be consistent with the results found in the literature. The following results showed that the best learning rate is 0.00002 with both the BERT and BioBERT models.

To complete the ablation study, several experiments with different learning rates were carried out using LSTM and BiLSTM models. The selected three learning rate values were chosen as they are the most used in literature. Adam optimizer, 50 training epochs, and 600 chuck size in the input were selected to match the results in the literature. Table 5 (for LSTM and BiLSTM) shows the calculated accuracies for different learning rates. Based on the results shown below, the learning rate was chosen to be 0.001, as it achieved the best results with LSTM model and an approximately good result (0.651) for the BiLSTM compared to the best value (0.652).

4. Conclusion

This research aims to introduce a multi-classification framework employing natural language processing methods. This framework categorizes genetic mutations by utilizing clinical evidence, specifically the textual explanations of these mutations. Thereby facilitating the advancement of individualized cancer therapy. In this study, NLP methods are utilized to construct a multi-label classifier. Text transformation models, specifically Glove, are employed to convert text into a matrix of token counts. The suggested framework is constructed using four deep learning classification models: BERT, BioBERT, Glove-based LSTM, and BiLSTM. The assessment utilizes a confusion matrix to evaluate the models' performance. Ultimately, empirical findings indicate that the BioBERT model outperformed the other suggested classifiers, achieving the highest accuracy of 86 % and F1-score of 87 % by optimizing the learning rate which greatly enhances the accuracy of the model. Furthermore, the suggested BioBERT greatly outperformed other models found in the literature. Moreover, the utilization of truncation for the input text enhances the results achieved by the BioBERT model. The results revealed that BioBERT outperforms BERT in the biomedical field because it is trained on biomedical data.

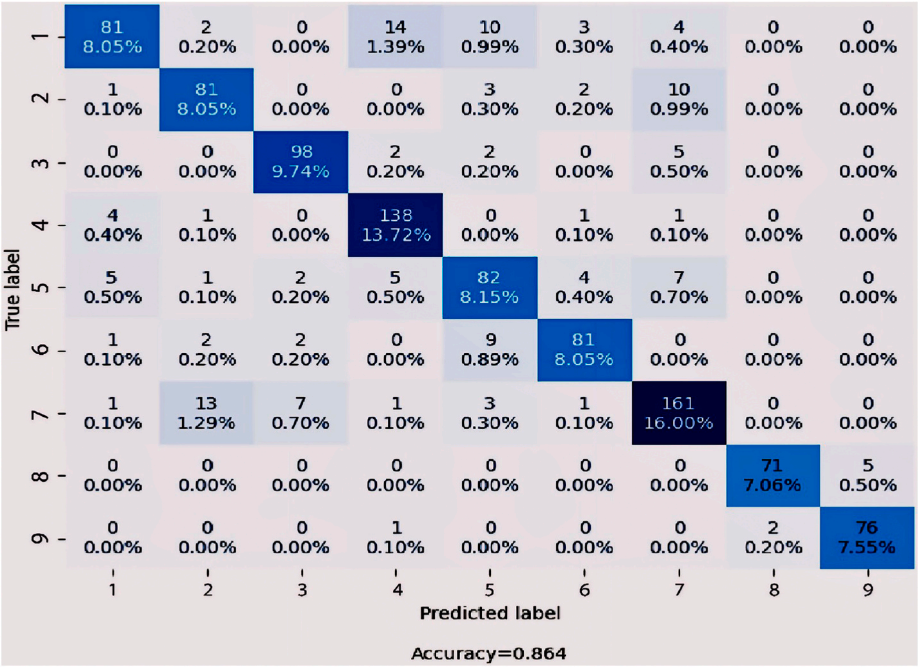


Fig. 8. Confusion Matrix for BioBERT using the Head + Tail truncation method.

Table 3
Comparison between proposed models against other models.

Model	Truncation	Word embedding	(Accuracy)
Deep-GenMut (BioBERT)	Head	–	0.780
Deep-GenMut (BioBERT)	Middle	–	0.717
Deep-GenMut (BioBERT)	**Head + Tail**	**-**	**0.864**
Deep-GenMut (BERT)	Head	–	0.701
Deep-GenMut (BERT)	Middle	–	0.703
Deep-GenMut (BERT)	Head + Tail	–	0.702
Deep-GenMut (LSTM)	–	GLOVE	0.643
Deep-GenMut (BiLSTM)	–	GLOVE	0.652
Logistic Regression [28]	–	TFIDF	0.385
Random Forest [28]	–	TFIDF	0.483
XGBoost [28]	–	TFIDF	0.497
Logistic Regression [28]	–	Word2Vec	0.467
Random Forest [28]	–	Word2Vec	0.450
XGBoost [28]	–	Word2Vec	0.482
RNN [28]	–	Pretrained Word2Vec	0.708
RNN [28]	–	Self-Trained Word2Vec	0.678
CNN [21]	–	TFIDF	0.648
Cascade neural network based on CNN and BiGRU [21]	–	TFIDF	0.811
Parallel hybrid neural network based on CNN and BiGRU [21]	–	TFIDF	0.844

Table 4
Learning rate impact on the accuracy of (BERT/BioBERT) models with Adam optimizer.

Learning rate	BioBERT (Accuracy)	BERT (Accuracy)
0.00002	**0.864**	**0.765**
0.00003	0.733	0.710
0.00004	0.807	0.708

5. Future work and limitations

One limitation of utilizing BioBERT in oncology mutation recognition stems from the size of the dataset employed for model training. The challenges inherent in the dataset, including the need to truncate text data due to the maximum input length limitation of

Table 5
Learning rate impact on the accuracy of (LSTM/BiLSTM) models with Adam Optimizer and 50 Epoch.

Learning rate	LSTM (Accuracy)	BiLSTM (Accuracy)
0.004	**0.643**	0.651
0.001	0.637	**0.652**
0.0001	0.527	0.534

512 tokens for BioBERT, may impact the classification process and potentially compromise classification accuracy. Future endeavors aimed at improving classification accuracy could explore novel approaches to address these limitations, such as augmenting the dataset with additional relevant samples or implementing advanced techniques to handle truncated data effectively. Moreover, future work could focus on leveraging the identified mutations and classes to predict drug-target interactions, thereby facilitating precision medicine initiatives in oncology and enhancing therapeutic strategies tailored to individual patients.

Data availability

The dataset explored in our study has been deposited into the publicly available repository Kaggle under the competition entitled "MSK redefining cancer treatment" (https://www.kaggle.com/c/msk-redefining-cancer-treatment/data).

CRediT authorship contribution statement

Emad A. Elsamahy: Writing – review & editing, Validation, Supervision, Methodology. **Asmaa E. Ahmed:** Writing – original draft, Visualization, Software, Resources, Methodology, Formal analysis. **Tahseen Shoala:** Validation, Supervision, Resources. **Fahima A. Maghraby:** Writing – review & editing, Validation, Supervision, Methodology.

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.

References

[1] Omar Ahmed, Adnan Brifcani, Gene expression classification based on deep learning, in: 4th Scientific International Conference Najaf, SICN 2019, 2019, https://doi.org/10.1109/SICN47020.2019.9019357.

[2] Peter D. Stenson, Matthew Mort, Edward V. Ball, Katy Evans, Matthew Hayden, Sally Heywood, Michelle Hussain, Andrew D. Phillips, David N. Cooper, The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies, Hum. Genet. (2017), https://doi.org/10.1007/s00439-017-1779-6. Human Genetics.

[3] K. BorkK. WulffL, Steinmüller-MaginI, BrænneP. Staubach-RenzG, WitzkeJ. Hardt, Hereditary angioedema with a mutation in the plasminogen gene, Allergy: European Journal of Allergy and Clinical Immunology (2018), https://doi.org/10.1111/all.13270.

[4] Takumi Onoyama, Shumpei Ishikawa, Hajime Isomoto, Gastric cancer and genomics: a review of literature, J Gastroenterol, ournal of Gastroenterology (2022), https://doi.org/10.1007/s00535-022-01879-3.

[5] N. Walters, L.T.H. Nguyen, J. Zhang, A. Shankaran, E. Reátegui, Extracellular vesicles as mediators of in vitro neutrophil swarming on a large-scale microparticle array, Lab Chip (2019), https://doi.org/10.1039/c9lc00483a.

[6] Ahmed Abd El-Hafeez Ibrahim, Atallah Ibrahim Hashad, Nigm El-Deen Mohamed Shawky, Aly Maher, " robust breast cancer diagnosis on four different datasets using multi-classifiers fusion", Int. J. Eng. Res. (2015), https://doi.org/10.17577/ijertv4is030173 and.

[7] AbdElhafeez Ibrahim Ahmed, Atallah Ibrahin Hashad, Eldin Mohamed Shawky Negm, A Comparison of Open-Source Data Mining Tools for Breast Cancer Classification, 2017, https://doi.org/10.4018/978-1-5225-2229-4.ch027.

[8] Yuhan Su, Hongxin Xiang, Haotian Xie, Yong Yu, Shiyan Dong, Zhaogang Yang, Na Zhao, Application of BERT to enable gene classification based on clinical evidence, BioMed Res. Int. (2020), https://doi.org/10.1155/2020/5491963.

[9] Carmen J. Allegra, R. Bryan Rumble, Stanley R. Hamilton, Pamela B. Mangu, Nancy Roach, Alexander Hantel, Richard L. Schilsky, Extended RAS gene mutation testing in metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy: American society of clinical oncology provisional clinical opinion update 2015, J. Clin. Oncol. (2016), https://doi.org/10.1200/JCO.2015.63.9674.

[10] Ahmad Ibrahim, Hoda K. Mohamed, Ali Maher, Baochang Zhang, A survey on human cancer categorization based on deep learning, Frontiers in Artificial Intelligence (2022), https://doi.org/10.3389/frai.2022.884749.

[11] Hai Hui Huang, Xiao Ying Liu, Yong Liang, Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2 + 2 regularization, PLoS One (2016), https://doi.org/10.1371/journal.pone.0149675.

[12] Qianqian Song, Mingyu Li, Qian Li, Xun Lu, Kun Song, Ziliang Zhang, Jiale Wei, Liang Zhang, Jiacheng Wei, Youqiong Ye, Jinyin Zha, Qiufen Zhang, Qiang Gao, Jiang Long, Xinyi Liu, Xuefeng Lu, Jian Zhang, DeepAlloDriver: a deep learning-based strategy to predict cancer driver mutations, Nucleic Acids Res. (2023), https://doi.org/10.1093/nar/gkad295.

[13] Yanxin Liu, Yifan Ma, Jingjing Zhang, Yuan Yuan, Jinqiao Wang, Exosomes: a novel therapeutic agent for cartilage and bone tissue regeneration, Dose Response (2019), https://doi.org/10.1177/1559325819892702.

[14] Peter D. Stenson, Matthew Mort, Edward V. Ball, Katy Howells, Andrew D. Phillips, Nick ST. Thomas, David N. Cooper, The human gene mutation database: 2008 update, Genome Med. (2009), https://doi.org/10.1186/gm13.

[15] Zihan Chen, Xingyu Li, Miaomiao Yang, Hong Zhang, Xu Steven Xu, Optimization of deep learning models for the prediction of gene mutations using unsupervised clustering, J. Pathol.: Clin. Res. (2023), https://doi.org/10.1002/cjp2.302.

[16] Arkadiusz Gertych, Zaneta Swiderska-Chada, Zhaoxuan Ma, Nathan Ing, Tomasz Markiewicz, Szczepan Cierniak, Hootan Salemi, Samuel Guzman, Ann E. Walts, Beatrice S. Knudsen, Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides, Sci. Rep. (2019), https://doi.org/10.1038/s41598-018-37638-9.

[17] Antonio R. Lucena-Araujo, Diego A. Pereira-Martins, Luisa C.A. Koury, Juan Luiz Coelho-Silva, Raul Antônio Morais Melo, Rosane Bittencourt, Katia B. Pagnano, Ricardo Pasquini, Carlos Chiattone, Evandro Maranhão Fagundes, Maria de Lourdes Chauffaille, Stanley L. Schrier, Martin S. Tallman, Raul C. Ribeiro,

David Grimwade, Arnold Ganser, Bob Lowenberg, Francesco Lo Coco, Miguel A. Sanz, Nancy Berliner, Eduardo Magalhães Rego, Combining gene mutation with gene expression analysis improves outcome prediction in acute promyelocytic leukemia, Blood (2019), https://doi.org/10.1182/blood.2019000239.

[18] Joyeeta DeyDhyani Desai, NLP based approach for classification of mental health issues using LSTM and GloVe embeddings, International Journal of Advanced Research in Science, Communication and Technology (2022), https://doi.org/10.48175/ijarsct-2296.

[19] Iker Huerga, Wendy Kan, Personalized Medicine: Redefining Cancer Treatment, Kaggle, 2017. https://kaggle.com/competitions/msk-redefining-cancer-treatment.

[20] Ralf C. Staudemeyer, Eric Rothstein Morris, Understanding LSTM– a tutorial into long short-term memory recurrent neural networks, arXiv preprint arXiv:1909.09586 (2019).

[21] Jingwen Xu, Xuling Zheng, Min Jiang, Gene mutation classification using CNN and BiGRU network, in: 9th International Conference on Information Science and Technology, ICIST 2019, 2019, https://doi.org/10.1109/ICIST.2019.8836846.

[22] Omar Einea, Ashraf Elnagar, Ridhwan Al Debsi, Sanad: single-label Arabic news articles dataset for automatic text categorization, Data Brief 25 (2019) 104076.

[23] Qian LiHao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, Lifang He, A survey on text classification: from traditional to deep learning, ACM Transactions on Intelligent Systems and Technology (2022), https://doi.org/10.1145/3495162.

[24] Spyros Makrida, kisEvangelos Spiliotis, Vassilios Assimakopoulos, Statistical and machine learning forecasting methods: concerns and ways forward, PLoS One (2018), https://doi.org/10.1371/journal.pone.0194889.

[25] Gangmin Li, Bei Yao, Classification of genetic mutations for cancer treatment with machine learning approaches, in: Conference Paper, 2018.

[26] Jingwen Xu, Xuling Zheng, Min Jiang, Gene mutation classification using CNN and BiGRU network, in: 9th International Conference on Information Science and Technology, ICIST 2019, 2019, https://doi.org/10.1109/ICIST.2019.8836846.

[27] Akash Kumar, Kandibanda Sai Santhosh, Personalized medicine: redefining cancer treatment using machine learning, International Journal of Engineering Applied Sciences and Technology (2020), https://doi.org/10.33564/ijeast.2020.v05i08.031.

[28] Meenu Gupta, Hao Wu, Simrann Arora, Akash Gupta, Gopal Chaudhary, Hua Qiaozhi, Gene mutation classification through text evidence facilitating cancer tumor detection, Journal of Healthcare Engineering (2021), https://doi.org/10.1155/2021/8689873.

[29] S. Aburass, O. Dorgham, J.A. Shaqsi, A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe, arXiv (Cornell University) (Jul. 2023), https://doi.org/10.48550/arxiv.2307.14361.

[30] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Jones A.N., Gomez A.N., Kaiser L., Polosukhin L., Attention is all you need, Adv. Neural Inf. Process. Syst. (2017),(Vol. 30).

[31] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, Chan Ho So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2020), https://doi.org/10.1093/bioinformatics/btz682.

[32] Winda Kurnia Sari, Dian Palupi Rini, Reza Firsandaya Malik, Text classification using long short-term memory with GloVe features, urnal Ilmiah Teknik Elektro Komputer dan Informatika (2020), https://doi.org/10.26555/jiteki.v5i2.15021.

[33] Marjan Kamyab, Guohua Liu, Michael Adjeisah, Attention-based CNN and Bi-LSTM model based on TF-IDF and GloVe word embedding for sentiment analysis, Appl. Sci. (2021), https://doi.org/10.3390/app112311255.

[34] Jeffrey Pennington, Richard Socher, Christopher Manning, "GloVe: global vectors for word representation", in: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, https://doi.org/10.3115/v1/d14-1162.

[35] Sudeb Das, Chowdhury Manish, Malay Kumar Kundu, " Brain MR. Image classification using multiscale geometric analysis of ripplet", Prog. Electromagn. Res. (2013), https://doi.org/10.2528/PIER13010105.

[36] Shang Gao, Mohammed Alawad, M. Todd Young, John Gounley, Noah Schaeffekoetterm, Hong Jun Yoon, Xiao Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, Linda Coyle, Georgia Tourassi, Limitations of transformers on clinical text classification, IEEE Journal of Biomedical and Health Informatics (2021), https://doi.org/10.1109/JBHI.2021.3062322.

[37] Munikoti Mahati, V. P. Subramanyam Rallabandi Srikantamurthy, Dawood Babu Dudekula, Sathishkumar Natarajan, Junhyung Park, Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning, BMC Med. Imag. (2023), https://doi.org/10.1186/s12880-023-00964-0.

[38] Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin, The performance of LSTM and BiLSTM in forecasting time series, in: Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019, 2019, https://doi.org/10.1109/BigData47090.2019.9005997.

[39] Devlin Jacob, Ming Wei Chang, Kenton Lee, Kristina Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.

[40] Alexander Turchin, Stanislav Masharsky, Marinka Zitnik, Comparison of BERT implementations for natural language processing of narrative medical documents, Inform. Med. Unlocked (2023), https://doi.org/10.1016/j.imu.2022.101139.

[41] Chandra Bhushana Rao Killi, Narayanan Balakrishnan, Chinta Someswara Rao, Classification of fake news using deep learning-based GloVE-LSTM model, International Journal of Safety and Security Engineering (2022), https://doi.org/10.18280/ijsse.120512.

[42] Arvind T. Mohan*, Datta V. Gaitonde, A deep learning based approach to reduced order modeling for turbulent flow control using LSTM neural networks, arXiv: 1804.09269v1 (2018) [physics.comp-ph] 24 Apr.