JURNAL INFOTEL
VOL. 17, NO. 2, MAY 2025, PP. 299–319. DOI:10.20895/INFOTEL.V17I2.1320

RESEARCH ARTICLE

Enhancing Disease Diagnosis Coding: A Deep Learning Approach with Bidirectional GRU For ICD-10 Classification

Aqge Priwibowo[1,*], Chandra Kusuma Dewa[2], and Ahmad Luthfi[3]

[1,2,3]Informatics Master Program, Faculty of Industrial Technology, Universitas Islam Indonesia, 55584, Indonesia

*Corresponding email: 22917025@students.uii.ac.id

Received: January 29, 2025; Revised: May 24, 2025; Accepted: May 26, 2025.

Abstract: The health insurance claim in hospitals involves selecting specific ICD-10 codes for primary diagnosis texts. With rising claim volumes, the need for faster, more accurate coding is critical. This study develops a deep learning model to classify diagnosis texts into relevant ICD-10 codes using 9,982 original medical records from a national referral hospital under the Indonesian Ministry of Health. The classification method employs a BiGRU layer architecture, known for its effectiveness in handling sequential data, such as diagnosis texts. BiGRU operates bidirectionally, enhancing the model's ability to capture the context from both past and future sequences. In this architecture, the BiGRU layer serves as the classification layer, stacked above the BERT layer, which functions as the vector embedding layer, converting text into numerical representations for the model. The results of the study demonstrate a promising solution for codifying primary diagnosis texts, achieving a precision of 82.18% and a recall of 81.59%. Despite the strong performance of the model, further improvements are possible. Interestingly, the study also observed that the size of the class volume per ICD-10 code is not the only factor affecting classification performance, as some classes with smaller volumes exhibited better classification results. However, merging rare classes did not improve performance and even worsened it, suggesting that better ways to handle underrepresented classes are needed. Experiments with different embedding layers, such as IndoBERT and BioClinicalBERT, and hyperparameter tuning yielded minimal performance gains, suggesting the need for alternative optimization strategies.

Keywords: BERT, BiGRU, ICD-10, Medical Text Classification, Stacked layer

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

The financial sustainability of healthcare facilities, such as hospitals, is closely related to the effectiveness of insurance claim management, which is crucial to maintaining stability [1]. A significant portion of hospital revenue is generated from insurance claims, including those from both the private and public sectors. To improve public welfare, the Indonesian government introduced the National Health Insurance program, BPJS Health [2]. Hospitals that collaborate with BPJS Health use a retrospective payment method, in which service costs are calculated based on case-based financing determined by the Ministry of Health [3], [4]. However, delays in the processing of claims can result in delayed payments, which can negatively impact hospital operations [5]. Therefore, processing claims promptly is crucial to maintain smoothness.

Accurate codification of diagnoses and medical procedures is crucial in grouping them appropriately within the prospective payment method in the healthcare system [6, 7]. Specifically, the accurate assignment of ICD-10 diagnosis codes plays a crucial role in determining the cost of healthcare services billed to patients [8]. This process is carried out by medical record officers, who carefully examine medical records and assign the relevant ICD-10 codes. The accuracy of this codification is vital for ensuring proper healthcare payments [8, 9]. Specifically, errors in coding, incorrect data input, and improper placement of diagnoses by medical record officers are then common causes of payment delays [2, 10]. As patient volume continues to rise, ensuring an adequate number of staff to handle codification accurately and efficiently has become increasingly crucial [9].

In this context, the primary focus is on addressing cases associated with the codification process conducted by the medical records team. More precisely, from a data science viewpoint, this codification process involves classifying diagnostic text into the relevant ICD-10 categories. The diagnostic text serves as a feature, with the ICD-10 code functioning as its classification label.

A major challenge in medical text classification is extracting meaningful information from unstructured medical descriptions, particularly those that utilize specialized medical terminology [11]. Medical data is often presented in formats that are difficult to interpret and analyze automatically, necessitating the development of specialized methods to process this text. Previous studies have explored various approaches to address this challenge. Another study [12] utilized a dual path model, combining a convolutional model to extract local features and a BiGRU pathway to capture long-term dependencies, which improved the accuracy of classification. In contrast, simpler models [13], such as those utilizing stacked layers with BERT embeddings, have demonstrated that leveraging enhanced word embeddings, like BERT, enables BiGRU to function more effectively. This approach has proven particularly successful in text classification tasks, including Chinese language processing, by providing more accurate contextual word representations and enhancing overall performance.

Based on the discussion, this study is framed around two main objectives. First, how can a model be constructed to effectively encode diagnostic text by classifying it into relevant ICD-10 code categories? Second, how accurate is the model in performing this codification? By addressing these two questions, this research aims to make a substantial contribution to improving the codification process in medical data management while simultaneously mitigating the potential for errors in health insurance claims.

1.1 Overview of Key Concepts: ICD-10, BERT, and BiGRU

To fully comprehend the diagnosis classification process in this research, it is crucial to understand the structure and purpose of the ICD-10 coding system. The ICD-10 coding system is divided into two components: the ICD-10 Clinical Modification for diagnoses and the ICD-10 Procedure Coding System (PCS) for procedures [14]. These codes consist of seven characters, a combination of numbers and letters. The first three characters indicate the type of diagnosis, the next two specify the particular diagnosis, and the final two describe the severity and other related details.

One of the most powerful embedding methods in natural language processing is BERT, recognized for its ability to generate contextual embeddings, even for out-of-vocabulary words [15]. In text classification, BERT is used to extract these embeddings by passing input sequences through a pre-trained model that understands the relationships between words in context. BERT-base, for instance, processes input sequences of up to 512 tokens, and the final hidden state is typically used as a representation of the entire sequence for classification tasks [16].

The Gated Recurrent Unit (GRU) is an evolution of the Recurrent Neural Network (RNN) architecture, designed for sequential data processing tasks, such as text, by using the output of the previous layer as input for the next [16]. GRU incorporates gating mechanisms, such as the forget gate to discard less important information and the remember gate to retain essential information [17]. A variation of GRU, the Bidirectional GRU (BiGRU), processes data in both forward and backward directions using independent parameters while sharing the same word vectors [18].

1.2 Related Work

Previous studies in medical text classification have made significant contributions using various techniques. Duarte [19] focused on classifying cause-of-death diagnoses into ICD-10 codes. His approach utilized word embeddings derived from death certificate texts, combined with recurrent units and neural attention mechanisms. This method achieved a precision of 68.52%, a recall of 63.78%, and an F1-score of 65.48%. Another study proposed a hierarchical LSTM architecture to map ICD-10 codes from clinical documents [20]. The goal was to address the complexity of the large ICD-10 code set and the limitations of available datasets. This approach showed promising results, with a precision and recall of 80.67%. Lastly, a study using Naive Bayes, Support Vector Machine, and Logistic Regression [21] categorized medical diagnoses into ICD-10 groups. Logistic regression achieved the highest precision, recall, and F1-score at 86%, 91%, and 88%, respectively.

While previous studies laid a strong foundation, our research stands out by using state-of-the-art transformer-based models, BERT, in combination with BiGRU for advanced feature extraction and sequence learning. This approach enables a deeper contextual understanding of medical texts. Moreover, while previous works have focused on various types of medical texts, our study utilizes a specialized dataset consisting of primary diagnostic texts from the National Hospital. This dataset provides a focused context for investigating the potential of advanced methods in diagnostic text classification. By combining the power of BERT with BiGRU, we aim to improve diagnostic text classification and achieve higher performance.

Prabhakar and Won developed a model using convolutional layers and BiGRU in parallel pathways for processing the Hallmarks and AIM datasets [12]. The architecture started

with a word embedding layer to convert text into contextually rich vectors, followed by CNN layers for extracting local features. The BiGRU pathway captured long-term dependencies in the text, with multi-head attention focusing on different parts of the text. The model achieved an accuracy of 95.76%, but the dual-channel design required significant computational resources as complexity increased [22].

Other research emphasized the effectiveness of simpler text classification methods that rely on stacked layers rather than segmented channels [13]. This study utilized BERT and BiGRU in conjunction for text classification, particularly in the context of Chinese text. Despite the simple architecture, BERT accurately represented word contexts, while BiGRU extracted bidirectional features. The approach achieved an accuracy of up to 95% across various text domains.

This simplified approach was successfully applied to medical text classification, particularly with specialized word embeddings, such as BioBERT [14]. Research has demonstrated that models using BioBERT (BGA and B2GA) improve accuracy by up to 38% compared to conventional BERT. Similarly, BiGRU-based models with self-attention efficiently extracted semantic features for classification [23]. These methods promise effective performance even with simpler architectures.

Our approach aligns with these previous studies by utilizing BERT in combination with BiGRU, focusing on a simpler stacked-layer architecture. This approach offers a balance between high performance and computational efficiency. By leveraging the strengths of BERT's contextual embeddings and BiGRU's ability to model long-term dependencies, we aim to enhance text classification in the medical text domain, specifically using diagnostic text datasets, without the computational overhead associated with more complex architectures.

2 Research Method

This study was carried out by adhering to the model development process stages as outlined in previous research [24], [25,26] as depicted in Figure 1.

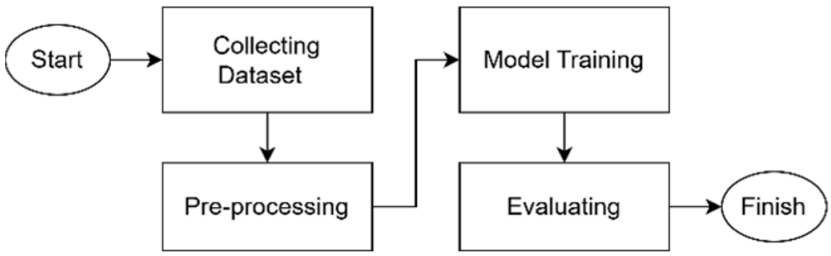

Figure 1: Flow of the model development process.

2.1 Dataset

This study utilized medical records from a national referral hospital's Information System (SIMRS), focusing on inpatient cases from 2021 to 2023. Initially, 10,000 rows of data were

collected, each comprising an ICD-10 code paired with the corresponding primary diagnosis text. During the extraction process, some rows were found to be corrupt or incomplete, resulting in unusable data. After thorough cleaning, 9,982 rows remained suitable for analysis. The dataset was exported into CSV or XLS formats, with preprocessing steps to address inconsistencies such as operational characters (e.g., '=' and '-'), empty fields, and other incompatible entries.

Table 1 illustrates several data features along with their corresponding class labels. It shows how some data points shift their labels due to label mapping. Several instances show inconsistencies in labeling, where similar features are assigned to different labels. These inconsistencies are observed across both small and large classes, indicating potential issues with label assignment and mapping. In addition, there are inconsistencies within the features themselves. These observations will be further explored in the discussion section.

For initial insights, the dataset was grouped to analyze the distribution of ICD-10 codes and diagnosis texts. Table 2 presents the results of grouping and sorting the dataset. This study focuses on ICD-10 groups with a minimum of 500 data rows, serving as class labels for classification. Other data groups are mapped into a special code labeled "NN" to simulate unrecognized codes. Table 3 displays the data population resulting from the mapping of ICD-10 codes into class labels. It shows that the NN label has the highest count, as it aggregates various ICD-10 codes into one category, as illustrated by the pie distribution in Figure 2. Subsequently, the cleaned dataset was randomly divided into three subsets: 70% for training, 20% for validation, and 10% for testing, ensuring a balanced and representative sample for model development.

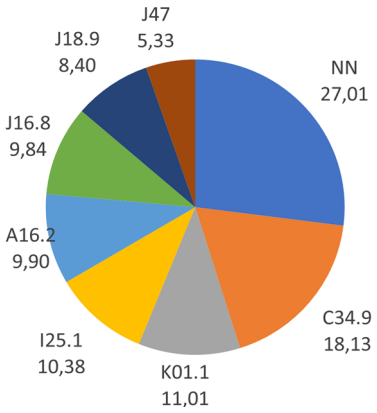

Figure 2: Distribution of mapped labels dataset.

2.2 Preprocessing

Text preprocessing is a crucial process that refines, cleans, and standardizes textual data to make it suitable for machine learning and deep learning systems. Common preprocessing techniques include text cleaning, tokenization, removal of irrelevant elements, conversion, correction, stopword removal, stemming, and lemmatization 15. The primary aim is to remove unnecessary content, resulting in a clean and structured text dataset.

Table 1: Illustration of The dataset with various labels and their corresponding mappings

ICD10	Mapped	Primary Diagnosis	Primary Diagnosis (Original)
A16.2	A16.2	1. Type 1 respiratory failure high-risk HAP 2. Septic shock ec HAP ...	1. Gagal napas tipe 1 ec HAP high risk 2. Syok sepsis ec HAP ...
C50.9	NN	1. Type 2 respiratory failure 2. Septic shock dd/hypovolemic 3. CAP with ...	1. Gagal napas tipe 2 2. Syok sepsis dd/hipovolemik 3. CAP dengan ...
E11.5	NN	1. Right foot gangrene post transtibial amputation POD 4 (29/9/23)	1. Gangren pedis dextra post amputasi transtibial dextra POD 4 (29/9/23)
E11.5	NN	1. Diabetic right foot gangrene pre-debridement with infection ...	1. Gangren pedis diabetikum dextra pro debridement dengan Infeksi ...
A09.9	NN	1. Acute gastroenteritis with moderate dehydration, resolved	1. Gastroenteritis akut dengan dehidrasi sedang teratasi
I63.8	NN	1. Right hemiparesis and motor aphasia with a history of ...	1. Hemiparesis dextra et afasia motorik dengan riwayat penurunan ...
J16.8	J16.8	1. Grade 4 hemorrhoids	1. Hemorrhoid grade 4
J16.8	J16.8	1. Pancytopenia history of severe neutropenia, suspected HIV infiltration	1. Pansitopenia riwayat neutropenia berat curiga infiltrasi HIV ...
K30	NN	1. Pancytopenia with neutropenia history, suspected HIV infiltration ...	1. Pansitopenia dengan riwayat neutropenia curiga infiltrasi HIV ...
I12.0	NN	1. Decreased consciousness ec sepsis 2. Upper GI bleeding ec suspected DIC ...	1. Penurunan kesadaran ec Sepsis 2. Upper GI bleeding ec Susp DIC...
J16.8	J16.8	1. Decreased consciousness ec septic shock due to CAP	1. Penurunan kesadaran ec syok sepsis ec CAP
J90	NN	CAP (104) Age, heart, hematocrit, pleural effusion	CAP (104) Usia, jantung, hematokrit, Efusi pleura
J16.8	J16.8	CAP (Community-Acquired Pneumonia)	CAP (Community Acquired Pneumonia)
A16.2	A16.2	CAP (Severe) PSI 99 (age, gender, respiratory rate, sodium)	CAP (Severe) PSI 99 (usia jenis kelamin, RR, natrium)

In this study, preprocessing was also implemented to minimize differences arising from editorial variations in representing identical data. This involved standardization techniques, addressing issues such as inconsistent use of Arabic and Roman numerals, irregular spacing, and normalization to unify the use of abbreviations and hierarchical references. Case folding was applied to convert all text to lowercase, ensuring uniformity across the dataset. The process was followed by tokenization and word vector embedding using BERT, which was incorporated into the model training process. The entire preprocessing pipeline was executed in a single iteration. While this approach posed the risk of requiring complete retraining if an error occurred, it eliminated the need for program code adjustments during implementation, streamlining the overall process.

Table 2: ICD-10 code frequency distribution

ICD-10	All Diagnosis	Unique Diagnosis	Label Map
C34.9	1792	1692	C34.9
A16.2	1088	785	A16.2
I25.1	1026	411	I25.1
J16.8	978	617	J16.8
K01.1	972	513	K01.1
J18.9	830	273	J18.9
J47	527	270	J47
C53.9	489	278	NN
C50.9	194	187	NN
A09.9	194	157	NN
I12.0	193	181	NN
C56	184	180	NN
E11.5	183	174	NN
I63.8	166	154	NN
E11.9	165	162	NN
J90	153	147	NN
I21.4	141	78	NN
A15.0	137	132	NN
K30	123	114	NN
J46	122	101	NN
Z03.1	71	69	NN
D38.1	65	64	NN
U07.1	47	44	NN
S06.0	42	36	NN
Total	9882	6819	24

Table 3: Distribution of ICD-10 codes in mapped labels for classification purposes

New ICD-10	All Data	Unique data
NN	2669	2258
C34.9	1792	1692
A16.2	1088	785
I25.1	1026	411
J16.8	978	617
K01.1	972	513
J18.9	830	273
J47	527	270
Total	9882	6819

2.3 Model Building

The proposed model in this study adopts frameworks from [16, 27–29]. The process begins by dividing the dataset into three parts: 70% for the training dataset, 20% for the validation dataset, and 10% for the final testing dataset. The deep learning architecture used to train the model consists of three main layers. The first layer is the BERT layer, which converts each word in the sentence into a vector using the BERT pretraining model [13, 27, 30]. The output vectors from BERT are then processed by the BiGRU layer, which extracts both

semantic and temporal features from the text. This is followed by a fully connected neural network classification layer, with the final SoftMax layer used to determine the category [27, 31]. Unlike other models that typically use the output from the BiGRU layer directly in the encoder or classifier, this proposed model focuses on the sequential integration of layers to enhance the effectiveness of text classification. (See Figure 3 for a schematic of the model architecture.) The training process is conducted using several sets of hyperparameters and variations in embedding for a small number of epochs to gain insights into the optimal setup. Subsequently, the training proceeds with a limit of 300 epochs, monitoring performance improvements and watching for signs of overfitting.

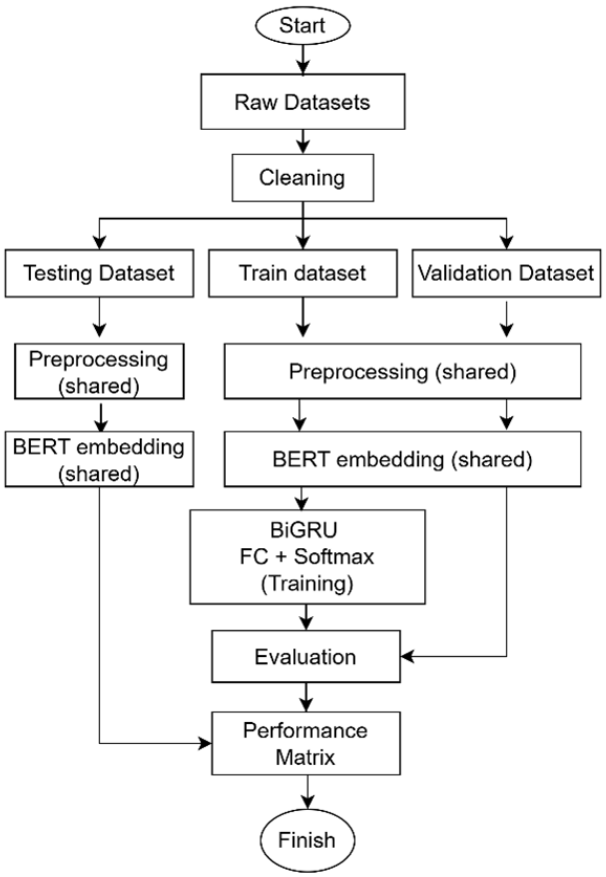

Figure 3: Schematic diagram of the model building process.

3 Results

3.1 Exploring Token Distribution

A sentence can be decomposed into a list of words, which can then be used to reconstruct the original sentence 15. This process is known as tokenization. In BERT, each token is converted into a numerical representation known as an embedding. Word embeddings are vectors that capture the semantic meaning of tokens, including positional and segment embeddings to represent the token sequence. The histogram in Figure 4 shows the distribution of the frequency of token length, representing the percentage of data that share the same number of tokens. Diagnostic tests are characterized by a high frequency of shorter tokens, as evidenced by their predominance of short sentences. This variation in sentence length, or data inconsistency, is important for analyzing performance metrics such as precision and recall during model evaluation stages.

To complement the analysis of the model, detailed information on the token distribution for each ICD-10 code group is also essential. Figure 5 sequentially displays histograms of token frequencies for each ICD-10 class label, providing insights into how tokens are distributed within specific categories. Some classes exhibit a more even distribution compared to others, which may indicate a balanced spread of sentence lengths. Notably, classes C34.9 and K01.1 show distributions resembling a normal distribution without skew, suggesting uniformity in token frequency that could impact model predictions differently. It has been observed that classes with a more balanced token-length distribution tend to perform better in classification tasks. This observation is further supported by the results shown in Figure 11, the confusion matrix, and the performance of the model in Table 7, both of which are presented in the following section.

3.2 Model Training

The model was trained using the Stacked BERT-BiGRU architecture with the parameters listed in Table 4, following several iterations of parameter adjustments tested during the study to achieve optimal performance. The training process involves calculating the loss, which measures the difference between the model's predictions and the actual values. For text classification, the training loss was computed using Cross Entropy Loss 30. During the first five epochs, a significant reduction in error was observed, followed by a continuous decrease in both training and validation losses. As shown in Figure 6, after epoch 36, the loss became minimal, with values approaching 10^{-6} per epoch for both training and validation.

After epoch 36, the performance improvement became less balanced. Figure 7 illustrates that the performance gains of the validation dataset were slower than those of the training dataset. The precision measure between the two datasets only differed by 0.0003 at epoch 35, but by epoch 300, it had grown to 0.0457. Although overfitting indicators have not fully emerged, the performance gap between the two datasets is noticeable. To anticipate overfitting, model states were manually backed up at five-epoch intervals before overfitting signs appeared. After signs of saturation were observed beyond epoch 40, further backups were conducted manually at irregular intervals, as no significant refinement in model performance was detected.

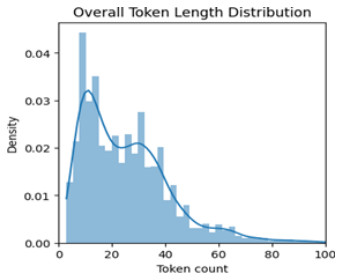

Figure 4: Histogram showing the distribution of token frequencies in diagnostic text data across the entire dataset.

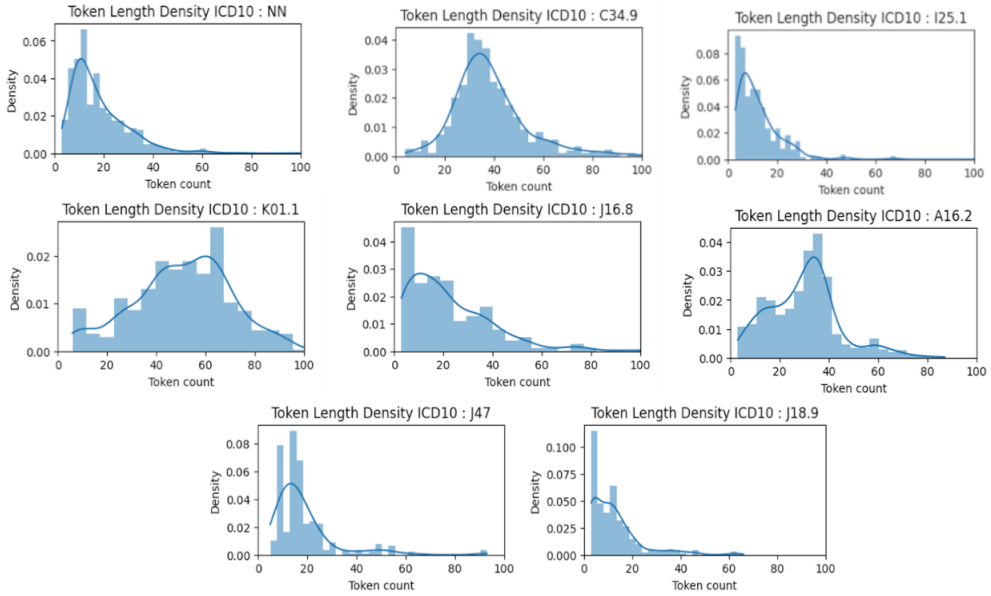

Figure 5: Token frequency distribution for each ICD-10 class label in the dataset.

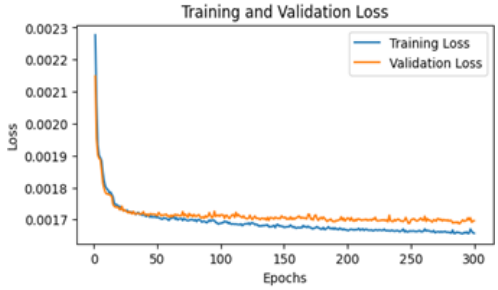

Figure 6: Reduction of training loss and validation loss over 300 epochs.

Table 4: Hyperparameters used for model training

Parameter	Value
BERT model	BERT Base
Learning rate	0.0001
Epoch	20
Dropout rate	0.2
Layer	2
Hidden layer dimension	128
Loss function	Cross Entropy
Optimizer	Adam

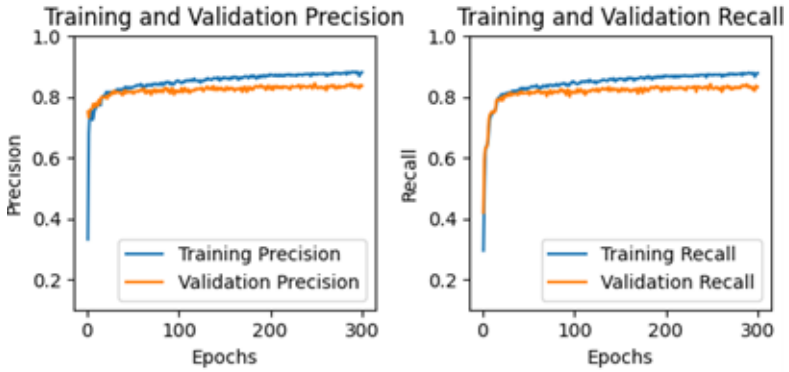

Figure 7: Comparing performance improvement for both training and validation datasets.

3.3 Hyperparameter Optimization

This study uses commonly used evaluation indices in text classification, namely precision (P), recall (R), and F1 score (F1) [30]. Precision indicates the accuracy of the model's predictions; recall represents the number of classes successfully identified by the model, and the F1 score is used when a balance between precision and recall is necessary. These three evaluation indices, as described by [12], are calculated using the following formulas:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

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

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

It is common practice to fine-tune various parameters when building a model to achieve optimal performance [29]. Research [13] identified the optimal parameters for the BERT-BiGRU classification as batch size 4, hidden layer size 768, and a learning rate of 1e-5. In this study, several hyperparameter combinations were explored to identify the optimal configuration for the text classification model. Specifically, experiments were conducted with various dropout rates ranging from 0.2 to 0.4 and different numbers of layers in BiGRU.

The results of these experiments are presented in Table 5, which illustrates the impact of hyperparameter changes on model performance. Up to epoch 15, it was identified that the optimal setup, as outlined in Table 5, was a dropout rate of 0.2 and two layers.

Table 5: Model performance evaluation with BERT-base model with learning rate of 1e-04 and variations in drop rate and number of nayers

Epoch	Evaluation	Dropout = 0.2		Dropout = 0.3		Dropout = 0.4	
		Layer = 2	Layer = 3	Layer = 2	Layer = 3	Layer = 2	Layer = 3
5	Precision	0.773	0.7583	0.7206	0.7377	0.7596	0.7514
	Recall	0.7113	0.672	0.6676	0.6631	0.6559	0.6863
	F1 score	0.6827	0.6105	0.6056	0.608	0.5914	0.6331
10	Precision	0.7761	0.7748	0.7636	0.7679	0.7669	0.7751
	Recall	0.7525	0.7498	0.7391	0.7462	0.7489	0.7551
	F1 score	0.7312	0.7308	0.7172	0.729	0.7296	0.7314
15	Precision	0.8031	0.8159	0.767	0.8146	0.7936	0.7822
	Recall	0.8034	0.798	0.7426	0.7971	0.7554	0.7819
	F1 score	0.8022	0.7969	0.7217	0.7776	0.7568	0.7793

3.4 Vector Embedding

Throughout the investigation, multiple-word embedding vectors were utilized to optimize the performance of the model in classification tasks and other related analyses [19]. As demonstrated by research [27], several word embeddings were employed in the study to explore their impact on model performance. In this context, experiments were conducted using three different pre-trained BERT models as embedding vectors. The performance of these pre-trained models BERT-base, Indo-BERT, and BioClinicalBERT - was compared to determine the most effective approach. The comparison results, as shown in Figure 8, indicate a similar upward performance trend for both training and validation across all three embedding vectors. Despite similar trends, Table 6 reveals that the embedding vector with the most significant impact in this experiment is BERT-Base, achieving a precision of 82.2% and a recall of 81.6% at epoch 20.

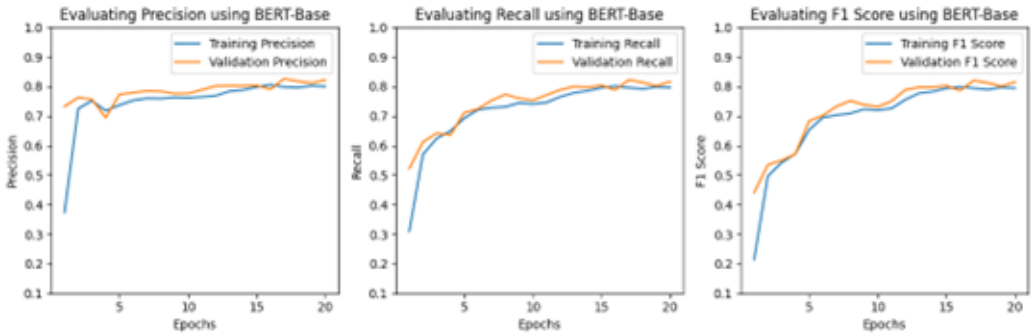

Figure 8: Model evaluation using BERT-base, Indo-BERT, and BioClinical-BERT.

Table 6: Model performance with variations in Vector Embedding

Epoch	BERT Based		Indo-BERT		BioClinical BERT	
	Precision	Recall	Precision	Recall	Precision	Recall
5	0.771	0.711	0.764	0.750	0.770	0.736
10	0.776	0.753	0.798	0.788	0.806	0.796
15	0.803	0.803	0.820	0.809	0.801	0.795
20	0.822	0.816	0.785	0.786	0.806	0.803

4 Discussion

4.1 Model Performance

Throughout the training process, continuous evaluations were performed on both the training and validation dataset to monitor performance improvements at each epoch, as practiced in [32]. These ongoing assessments allowed for real-time adjustments and provided valuable feedback on the model's progress. After training the model for 300 epochs, a final evaluation of the test dataset revealed strong performance, with a precision of 83.67%, a recall of 82.74%, an F1 score of 82.66%, and an accuracy of 81.92%. The high precision indicates that the model accurately identifies positive cases, with 83.04% of its positive predictions being correct. The recall of 82.74% reflects the model's capability to detect most of the actual positive cases. Despite these strong metrics, the accuracy of 81.92% indicates that there is still room for improvement in the overall correctness of the model.

Performance on both the training and validation datasets showed consistent improvement across the 300 epochs, as illustrated in Figure 9 and Figure 10. This steady upward trend highlights the model's effective learning capability. However, it is important to remain cautious about potential signs of overfitting. As depicted in Figure 7, there is a noticeable divergence between training and validation performance. While the validation dataset does not show a significant decline, the continuous increase in the training dataset's performance suggests that the model might be overly fitting to the training data. To address this issue, consider limiting the training to the epoch where the issues first arise, thereby preventing overfitting while still allowing for model improvement.

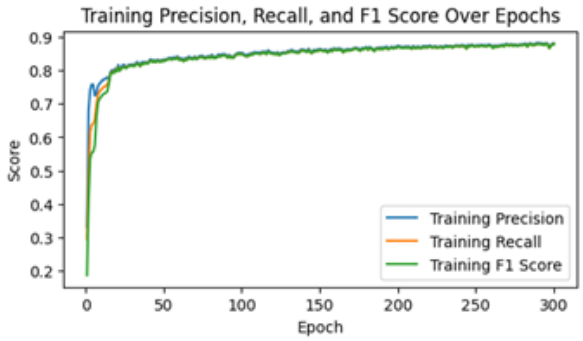

Figure 9: Performance gain on the training dataset: precision, recall, and F1 Score.

The confusion matrix for the test dataset offers a comprehensive view of the model's classification accuracy across different classes [33], [34]. The confusion matrix in Figure 11 provides a detailed view of the model's ability to predict each class, helping to identify specific classes where the model may struggle. This evaluation highlights overall performance and reveals class-level variations, indicating where the model performs well and where further refinement is necessary for more accurate class-specific predictions. By analyzing the matrix, as shown in Figure 11, the performance of each class/label can be evaluated, pinpointing the strengths and weaknesses of the model.

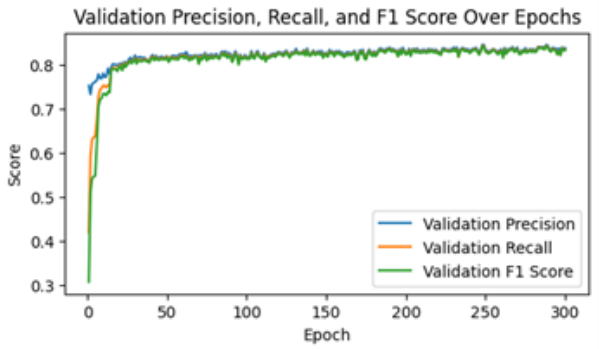

Figure 10: Performance gain on the validation dataset: precision, recall, and F1 Score.

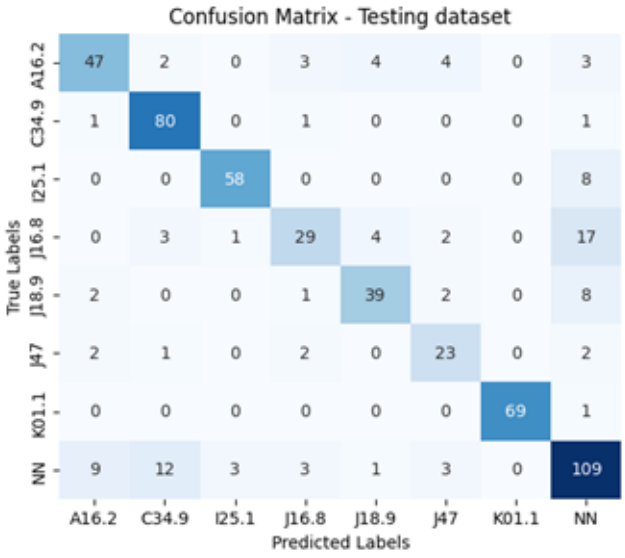

Figure 11: Confusion matrix, showing the model performance across all classes.

In general, larger datasets improve classification performance by providing more representative samples, reducing overfitting, and enhancing the model's ability to generalize to new data, as demonstrated by recent research [35, 36]. The strong correlation between

the available data population per class and the model's performance in predicting each class can be confirmed by revisiting Table 3. The classes with the most data population, except the NN class, show the highest performance. However, it is essential to consider the available data population not only in terms of raw data count but also in terms of unique data points. Duplicating data in a class can make the data size seem larger but could potentially lead to an overestimation of the model's training input [37]. The true impact of data availability on model performance should consider the diversity of data within each class, as repeated data points may not contribute effectively to the model's learning process.

However, the performance of the model is not solely influenced by the population size of a data set. Research [37] highlights that these six dimensions of data quality, consistency, completeness, accuracy of characteristics, uniqueness, accuracy of targets, and balance of target classes play a crucial role in determining model performance. In our study, we observed that some classes with smaller populations outperformed others with larger populations. This discrepancy can be attributed to the higher data quality in smaller classes. In Table 7, classes NN and J16.8 exhibit lower performance compared to other classes with smaller dataset sizes.

Table 7: Evaluation of model performance by class

Class	Precision (%)	Recall (%)	F1 Score (%)
K01.1	1.0000	0.9857	0.9928
I25.1	0.9206	0.8788	0.8992
C34.9	0.8163	0.9639	0.8831
NN	0.7315	0.7786	0.7543
A16.2	0.7705	0.7460	0.7580
J47	0.7188	0.7667	0.7420
J18.9	0.7500	0.7222	0.7358
J16.8	0.7632	0.5179	0.6170

Analysis of dimension completeness and feature accuracy requires domain-specific knowledge, particularly in the medical field. Completeness refers to the presence of all relevant and necessary data, while feature accuracy indicates how closely dataset values reflect the true values. However, upon further examination of Table 1 in the dataset section, it becomes apparent that both classes suffer from issues related to target accuracy. Class J16.8, in particular, is characterized by frequent mislabeling, with instances of multiple different labels assigned to the same samples. This inconsistency in the labeling contributes to inaccuracies in the ability of the model to accurately identify the class [38], thereby affecting its overall performance. Moreover, the features associated with Class J16.8 also demonstrate inconsistency, which hinders the model's capacity to learn from the data effectively.

NN class, on the other hand, is a composite class constructed from 17 smaller subclasses, and Table 1 reveals that it contains numerous instances of label duplication and feature overlap with other classes. These issues introduce significant noise, further complicating the classification process. According to research [37], such problems in data quality - specifically in terms of consistency, target accuracy, and class balance - are expected to degrade model performance. Even though Class NN contains a larger number of unique data points, the lack of quality in terms of data consistency and feature uniqueness leads to diminished predictive accuracy, confirming that data quality dimensions outweigh the mere size of the dataset.

Further investigation reveals a potential relationship between class performance and the distribution of tokens per class. While existing research, such as [39], suggests that token frequency may influence classification model performance and [40] discusses the impact of token length on model outcomes, research [41] demonstrates that a dataset with a normal distribution of token length frequencies and less skew correlates with higher performance. This study makes similar observations based on the data, as Figure 5 indicates that classes with a more continuous and evenly distributed token frequency tend to perform better.

This finding implies that not only the dataset size but also a balanced token representation within the classes may positively influence the model's accuracy in predicting those classes. For instance, classes such as K01.1, I25.1, and C34.9 exhibit token distributions that approach a normal distribution, which correlates with their higher performance. In contrast, while class NN exhibits a better token frequency distribution, as mentioned earlier, it is plagued by significant noise and inconsistency in its features, which ultimately hinder its performance despite the favorable token distribution.

4.2 Future Work

Recent findings have highlighted that the quality of data labeling has a significant impact on model performance, particularly in medical classification tasks. A recent study [37] highlights that inconsistencies, low accuracy, and poor class balance in labeled datasets significantly degrade the effectiveness of machine learning systems. These observations align with the study's findings, particularly regarding the NN and J16.8 classes, where issues such as duplicate entries, ambiguous label mappings, and feature overlap introduce considerable noise. Such problems suggest that the current annotation quality in the dataset requires systematic improvement to achieve better results in clinical environments.

To mitigate these issues, future work should focus on improving the annotation process. This includes involving multiple medical experts to cross-validate labels, refining clinical coding guidelines, and applying semi-automated quality control techniques to detect labeling inconsistencies [42]. This research shows that targeted data augmentation strategies can also enhance model robustness under noisy conditions, especially in medical NER tasks. Additionally, semi-supervised learning (SSL) approaches offer the potential to enhance text classification performance when labeled data are limited. SSL techniques leverage both labeled and unlabeled data to improve the accuracy of the model. Recent studies [43] have explored various SSL methods, including pseudo-labeling and self-training, to address challenges in text classification tasks. Another promising direction is the use of label correlation modeling. Research [44] introduced the LCFM method, which learns to infer missing or imprecise labels by exploiting inter-label relationships. Incorporating such approaches can enhance classification performance, particularly in settings where obtaining complete or clean annotations is challenging.

Finally, while data augmentation and semi-supervised learning techniques offer promising avenues for enhancing model performance, their application must be approached with caution. This study uses authentic diagnostic texts sourced from a national referral hospital, which reflect the linguistic and contextual realities of local clinical practice. Any additional methods must therefore be carefully evaluated to ensure that they do not introduce artifacts or biases that compromise the original characteristics of the data.

Preserving the integrity of real-world datasets is essential to maintain the clinical relevance and trustworthiness of deep learning models in healthcare settings.

4.3 Contribution

Although the BERT-BiGRU stacked architecture employed in this study is based on well-established methods in natural language processing, the value of this work lies in its real-world application. Specifically, it uses authentic diagnostic texts collected from a national referral hospital in Indonesia, capturing the local clinical language and context that are rarely represented in publicly available datasets. This practical implementation in a high-stakes environment such as insurance claim processing demonstrates the relevance and utility of such models beyond theoretical performance, highlighting the uniqueness of the dataset and the implementation setting. Beyond architectural design, the most significant contribution of this study lies in the development of a domain-specific trained model using authentic diagnosis texts from a national referral hospital. This model provides a practical, ready-to-implement solution for supporting ICD-10 classification in hospital business processes, particularly in contexts where accuracy of codification and operational efficiency are crucial.

5 Conclusion

This research aimed to develop a model capable of encoding diagnostic texts and evaluating its accuracy and effectiveness. Based on the evaluations conducted, the developed model demonstrated promising results, achieving an accuracy of 81.92%, a precision of 82.18%, and a recall of 81.59%. These figures indicate that the model can effectively encode diagnostic texts, thereby enhancing the accuracy of diagnostic coding and improving the health insurance claims process. Furthermore, the evaluation of the model revealed that the data volume per class influences its classification performance. Generally, a larger data population enhances predictive capability. Additionally, classes with a continuous and evenly distributed token frequency demonstrated better performance, suggesting that balanced token representation may hold promise for further exploration. In contrast, some classes performed poorly despite meeting the criteria for data volume and token distribution. These classes consist of merged classes, which lack distinct features and are inconsistent in their feature representations, making them difficult for the model to recognize. Duplication and inconsistent labeling were observed within these classes, likely contributing to their degraded performance. With the current model achieving an accuracy of 81%, this research presents opportunities for further improvement. Based on the findings, greater attention should be given to increasing data volume, improving the accuracy of labeling, and maintaining clear, consistent, and distinctive features in the dataset. These three factors have proven to be crucial in enhancing the predictive accuracy of the classification model. In addition to refining label quality, as outlined, future work may also explore data augmentation and semi-supervised learning, provided that these techniques are applied cautiously to avoid introducing bias or compromising the integrity of the original, real-world local dataset.

Acknowledgments

The researcher would like to express gratitude to the Director of Medical Services and the SIM RS Division at RSUD Persahabatan Jakarta for their invaluable support in supplying the necessary information and dataset for this research.

References

[1] D. Cavmak, "A study on the internal determinants of financial sustainability performance in private hospitals," *Social Science Research Network*, Apr. 2024. Rochester, NY: 4816229.

[2] I. A. Udin, "Factors causing delayed claims at the hospital in collaboration with health social security agency branch office of tasikmalaya," in *Proceeding of International Conference Sustainable Competitive Advantage*, vol. 3, Nov. 2022.

[3] S. Sulastri, M. Menap, and S. Sastrawan, "Problematic bpjs health patient pending claims in the implementation of ina-cbgs and management handling," *Jurnal Mantik*, vol. 7, Nov. 2023.

[4] Swandayana, "Analysis of the difference between ina-cbg rates and hospital rates for outpatient and inpatient services at fkrtl provider bpjs kesehatan mataram city," *Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram*, 2025. Accessed: Feb. 01, 2025.

[5] N. Putri, R. Semiarty, and N. A. Syah, "Health insurance (bpjs-kesehatan) late payment for hospital inpatient claims - a case study in west sumatra," *Berita Kedokteran Masyarakat*, vol. 36, Dec. 2020.

[6] Kementerian Kesehatan Republik Indonesia, "Peraturan menteri kesehatan republik indonesia nomor 52 tahun 2016 tentang standar tarif pelayanan kesehatan dalam penyelenggaraan program jaminan kesehatan," 2016. Regulation of the Minister of Health of the Republic of Indonesia.

[7] A. Susanto, "Reduced hospital revenue due to error code diagnosis in the implementation of ina-cbgs," *International Journal of Public Health Science (IJPHS)*, 2021. Accessed: Feb. 01, 2025.

[8] I. Saputra, S. M. Aljunid, and A. M. Nur, "The impact of casemix reimbursement on hospital revenue in indonesia," *Jurnal Sains Kesihatan Malaysia*, vol. 18, June 2020. Accessed: Feb. 01, 2025.

[9] N. Maimun, "Factors delayed of insurance claim service process (bpjs) at annisa pekanbaru maternity hospital," *Jurnal Kesehatan Komunitas*.

[10] M. Yastori, "Cases of dispute and pending claims in hospitals in the era of national health insurance," *icmr*, vol. 2, pp. 32–38, Jan. 2022.

[11] A. Chraibi *et al.*, "A deep learning framework for automated icd-10 coding," in *Public Health and Informatics*, pp. 347–351, IOS Press, 2021.

[12] S. K. Prabhakar and D.-O. Won, "Medical text classification using hybrid deep learning models with multihead attention," *Computational Intelligence and Neuroscience*, vol. 2021, p. e9425655, Sept. 2021.

[13] Q. Yu, Z. Wang, and K. Jiang, "Research on text classification based on bert-bigru model," in *Journal of Physics: Conference Series*, vol. 1746, p. 012019, Jan. 2021.

[14] W. Chen, F. Fang, P. Wang, J. Kan, W. Li, and W. Wu, "Research on medical text classification based on biobert-gru-attention," in *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pp. 213–219, Aug. 2022.

[15] D. Sarkar, *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Berkeley, CA: Apress, 2019.

[16] S. Shreyashree, P. Sunagar, S. Rajarajeswari, and A. Kanavalli, "Bert-based hybrid rnn model for multi-class text classification to study the effect of pre-trained word embeddings," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, 2022.

[17] E. Moons, A. Khanna, A. Akkasi, and M.-F. Moens, "A comparison of deep learning methods for icd coding of clinical records," *Applied Sciences*, vol. 10, no. 15, 2020.

[18] Y. Tian, "Multi-label text classification combining bert and bi-gru based on the attention mechanism," *Journal of Network Intelligence*, 2023.

[19] F. Duarte, B. Martins, C. S. Pinto, and M. J. Silva, "Deep neural models for icd-10 coding of death certificates and autopsy reports in free-text," *Journal of Biomedical Informatics*, vol. 80, pp. 64–77, Apr. 2018.

[20] S. S. Azam, M. Raju, V. Pagidimarri, and V. C. Kasivajjala, "Cascadenet: An lstm based deep learning model for automated icd-10 coding," in *Advances in Information and Communication* (K. Arai and R. Bhatia, eds.), Lecture Notes in Networks and Systems, pp. 55–74, Springer International Publishing, 2020.

[21] Z. A. Amin, W. Cholil, M. I. Herdiansyah, and E. S. Negara, "Analisa rekam medis elektronik untuk menentukan diagnosa medis dalam kategori bab icd 10 menggunakan machine learning," *PJSTI*, vol. 7, no. 2, pp. 127–132, 2021.

[22] C. Chen *et al.*, "Deep learning on computational-resource-limited platforms: A survey," *Mobile Information Systems*, vol. 2020, p. e8454327, Mar. 2020.

[23] T. Jiang and Z. Wang, "Text classification using bigru with directional self-attention," in *2022 11th International Conference of Information and Communication Technology (ICTech)*, pp. 394–397, Feb. 2022.

[24] A. Hanifa, Y. I. Kurniawan, J. H. Husen, A. K. Nugroho, and I. Permadi, "Prediction of patient length of stay using random forest method based on the indonesian national health insurance," *JURNAL INFOTEL*, vol. 15, Aug. 2023.

[25] S. Puspasari, R. Gustriansyah, and A. Sanmorino, "Forecasting a museum visit post pandemic using exponential smoothing model," *JURNAL INFOTEL*, vol. 15, Nov. 2023.

[26] A. Muis, S. Sunardi, and A. Yudhana, "Medical image classification of brain tumor using convolutional neural network algorithm," *JURNAL INFOTEL*, vol. 15, Aug. 2023.

[27] P. Ni, G. Li, P. C. K. Hung, and V. Chang, "Staresgru-cnn with cmedlms: A stacked residual gru-cnn with pre-trained biomedical language models for predictive intelligence," *Applied Soft Computing*, vol. 113, p. 107975, Dec. 2021.

[28] X. Zhang, Z. Wu, K. Liu, Z. Zhao, J. Wang, and C. Wu, "Text sentiment classification based on bert embedding and sliced multi-head self-attention bi-gru," *Sensors*, vol. 23, Jan. 2023.

[29] M. A. Parwez, M. Fazil, M. Arif, M. T. Nafis, and M. R. Auwul, "Biomedical text classification using augmented word representation based on distributional and relational contexts," *Computational Intelligence and Neuroscience*, vol. 2023, pp. 1–22, Feb. 2023.

[30] Q. Qin, S. Zhao, and C. Liu, "A bert-bigru-crf model for entity recognition of chinese electronic medical records," *Complexity*, vol. 2021, p. e6631837, Jan. 2021.

[31] S. Ouyang, Y. Shao, J. Du, and A. Li, "Scientific and technological text knowledge extraction method of based on word mixing and gru," *arXiv*, Mar. 2022.

[32] "View of banana and orange classification detection using convolutional neural network," 2025. Accessed: Jan. 11, 2025.

[33] S. S. Berutu, H. Budiati, J. Jatmika, and F. Gulo, "Data preprocessing approach for machine learning-based sentiment classification," *JURNAL INFOTEL*, vol. 15, Nov. 2023.

[34] A. Desiani *et al.*, "Weighted voting ensemble learning of cnn architectures for diabetic retinopathy classification," *JURNAL INFOTEL*, vol. 16, Feb. 2024.

[35] A. Althnian *et al.*, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Applied Sciences*, vol. 11, Jan. 2021.

[36] M. Arhami, F. Y. R. F, H. Hendrawaty, and A. Adriana, "A semantic segmentation of nucleus and cytoplasm in pap-smear images using modified u-net architecture," *JURNAL INFOTEL*, vol. 16, May 2024.

[37] S. Mohammed *et al.*, "The effects of data quality on machine learning performance," *arXiv*, Dec. 2024.

[38] V. Shah, T. Parashos, and A. Kumar, "How do categorical duplicates affect ml? a new benchmark and empirical analyses," *Proc. VLDB Endow.*, vol. 17, pp. 1391–1404, May 2024.

[39] O. Goldman, A. Caciularu, M. Eyal, K. Cao, I. Szpektor, and R. Tsarfaty, "Unpacking tokenization: Evaluating text compression and its correlation with model performance," *arXiv*, June 2024.

[40] "Sentiment analysis of comment texts based on bilstm," *IEEE Xplore*, 2019. Accessed: Jan. 18, 2025.

[41] S. Yehezkel and Y. Pinter, "Incorporating context into subword vocabularies," *arXiv*, Feb. 2023.

[42] H. Chen, L. Dan, Y. Lu, M. Chen, and J. Zhang, "An improved data augmentation approach and its application in medical named entity recognition," *BMC Medical Informatics and Decision Making*, vol. 24, p. 221, Aug. 2024.

[43] S. Al-Azzawi, G. Kovács, F. Nilsson, T. Adewumi, and M. Liwicki, "Nlp-ltu at semeval-2023 task 10: The impact of data augmentation and semi-supervised learning techniques on text classification performance on an imbalanced dataset," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (A. K. Ojha, A. S. Doğruöz, G. D. S. Martino, H. T. Madabushi, R. Kumar, and E. Sartori, eds.), pp. 1421–1427, Association for Computational Linguistics, July 2023.

[44] Y. Yu, Z. Zhou, X. Zheng, J. Gou, W. Ou, and F. Yuan, "Enhancing label correlations in multi-label classification through global-local label specific feature learning to fill missing labels," *Computers and Electrical Engineering*, vol. 113, p. 109037, Jan. 2024.