







## Article

# Hybrid Summarization of Medical Records for Predicting Length of Stay in the Intensive Care Unit

Soukaina Rhazzafe <sup>1,\*</sup>, Fabio Caraffini <sup>2</sup>, Simon Colreavy-Donnelly <sup>1</sup>, Younes Dhassi <sup>3</sup>, Stefan Kuhn <sup>4</sup>  
and Nikola S. Nikolov <sup>1</sup>

<sup>1</sup> Department of Computer Science and Information Systems, University of Limerick, V94 T9PX Limerick, Ireland; simon.colreavy@ul.ie (S.C.-D.); nikola.nikolov@ul.ie (N.S.N.)

<sup>2</sup> Department of Computer Science, Swansea University, Swansea SA2 8PP, UK; fabio.caraffini@swansea.ac.uk

<sup>3</sup> Sciences and Technology Faculty, Sidi Mohamed Ben Abdellah University, Fes 30050, Morocco; younes.dhassi@usmba.ac.ma

<sup>4</sup> Institute of Computer Science, Tartu University, 51009 Tartu, Estonia; stefan.kuhn@ut.ee

\* Correspondence: soukaina.rhazzafe@ul.ie

**Abstract:** Electronic health records (EHRs) are a critical tool in healthcare and capture a wide array of patient information that can inform clinical decision-making. However, the sheer volume and complexity of EHR data present challenges for healthcare providers, particularly in fast-paced environments such as intensive care units (ICUs). To address this problem, the automatic summarization of the main problems of patients from daily progress notes can be extremely helpful. Furthermore, by accurately predicting ICU patients' lengths of stay (LOSs), resource allocation and management can be optimized, allowing for a more efficient flow of patients within the healthcare system. This work proposes a hybrid method to summarize EHR notes and studies the potential of these summaries together with structured data for the prediction of LOSs of ICU patients. Our investigation demonstrates the effectiveness of combining extractive and abstractive summarization techniques with a concept-based method combined with a text-to-text transfer transformer (T5), which shows the most promising results. By integrating the generated summaries and diagnoses with other features, our study contributes to the accurate prediction of LOSs, with a support vector machine emerging as our best-performing classifier with an accuracy of 77.5%, surpassing existing systems and highlighting the potential for optimal allocation of resources within ICUs.

**Keywords:** natural language processing (NLP); text summarization; electronic health records (EHR); intensive care unit (ICU); length of stay (LOS); MIMIC-III; classification



**Citation:** Rhazzafe, S.; Caraffini, F.; Colreavy-Donnelly, S.; Dhassi, Y.; Kuhn, S.; Nikolov, N.S. Hybrid Summarization of Medical Records for Predicting Length of Stay in the Intensive Care Unit. *Appl. Sci.* **2024**, *14*, 5809. <https://doi.org/10.3390/app14135809>

Academic Editors: Antonio Pagliaro and Pierluca Sangiorgi

Received: 10 June 2024

Revised: 28 June 2024

Accepted: 1 July 2024

Published: 3 July 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The length of stay (LOS) for a hospitalized patient refers to the duration of their single admission measured in days. Besides serving as a key metric for assessing hospital resource utilization and the operational efficiency of healthcare systems, the LOS offers valuable insights into patient flow within care units [1].

In recent decades, hospitals worldwide have seen a notable decline in the average LOS, with the United States experiencing a reduction from approximately 20.5 days in 1960 to merely 6.5 days in 2021, making it among the shortest LOS globally [2]. However, despite this reduction, the cost of inpatient care remains exceedingly high. For instance, the average total ICU cost in the United States was USD 13,443 in 2021 [3].

While reducing the LOS may not always be feasible as a cost-saving measure, hospitals can still make savings by optimizing the allocation of personnel and resources. Accurately predicting the LOS can facilitate this optimization and serves as an effective strategy for healthcare services to implement preventative measures aimed at avoiding LOS extensions [1]. This is beneficial from different aspects, including the patient's condition and medical plan, family, the hospital, and the insurance company [4]. Additionally, the average

expenses associated with patients who experience a prolonged stay in the ICU are seven times higher compared to those who have a regular LOS [5], with a prolonged LOS being more than a week. Comorbidities, patient-specific factors such as age and frailty, and logistical challenges like delayed admissions and bed shortages all contribute to prolonged stays [6,7].

Furthermore, in the ICU, healthcare providers face the challenge of dealing with a huge amount of information and cognitive demands, leading to potential information overload and difficulties with processing critical patient details, which increases the risk of missed diagnoses and medical errors, impacting patient outcomes significantly [8].

To address these issues, our proposed solution is to automatically generate diagnoses and problems from progress notes, which, along with other clinical and demographic features, will predict ICU patients' LOSs. This automation helps reduce information overload and cognitive biases that may arise during the interpretation of complex patient data, aiding with understanding patient conditions accurately. In addition to that, by accurately predicting ICU patients' LOSs, resource allocation and management can be optimized through proactive planning of patient transfers and discharges.

Accordingly, our approach begins by tackling a challenge proposed by Physionet (PhysioNet: *The Research Resource for Complex Physiologic Signals*, 2024, <https://physionet.org> (accessed 10 June 2024)) ("BioNLP Workshop 2023 Shared Task 1A: Problem List Summarization", 2023, <https://physionet.org/content/bionlp-workshop-2023-task-1a/2.0.0> (accessed 10 June 2024)), which is a renowned platform supporting physiological signal research that is managed by MIT's Computational Physiology laboratory and backed by the National Institute of Biomedical Imaging and Bioengineering [9]. The challenge is entitled: "BioNLP Workshop 2023: Problem List Summarization" [10], and the main task is to generate a list of diagnoses and problems in a patient's daily care plan using input from the provider's progress notes during hospitalization in the ICU. The method then focuses on the extraction of additional features from the dataset that will contribute to a comprehensive understanding of each patient's medical profile. Finally, the generated problems and diagnoses and the extracted variables are used as features to predict the patient's LOS in the ICU.

We propose the first hybrid summarization approach that incorporates vital signs and laboratory results, with the goal of predicting the ICU LOS. While our summarization models do not outperform existing methods, our proposed ICU LOS classifier sets a new benchmark on the MIMIC-III dataset with an accuracy of 77.5%, significantly improving over previous methods and demonstrating the efficacy of our approach.

## 2. Background

This section provides an overview of both clinical text summarization and ICU LOS prediction; we outline the state-of-the-art methods and review published works in both domains.

### 2.1. Clinical Text Summarization

Text summarization is a process that produces a concise, fluent, and short summary of a longer document. Automatic text summarization systems are designed by applying one of the following general text summarization approaches [11].

#### 2.1.1. Extractive Text Summarization

Extractive text summarization involves selecting a few relevant sentences from the original document to create a concise summary based on their relevance, importance, or saliency to the overall meaning of the text. Extractive summarization methods can be categorized into various approaches, each with distinct techniques and objectives [11,12]. Concept-based extractive summarization methods focus on extracting concepts from the text using external knowledge bases to calculate the importance of the sentence. Graph-based methods use sentence-based graphs for document representation and to rank sentences. Topic-based summarization methods identify the main subject of the document using techniques like TF-IDF. Clustering-based extractive summarization methods aim to

identify the most central and important sentences in a cluster, as they include the important information related to the main subject; one well-known summarizer based on this method is BERTSummarizer, which employs the BERT (bidirectional-encoder transformer) [13] model to encode the input text and generate sentence embeddings. Despite offering speed and simplicity, extractive summarization can lead to redundancies and longer sentences, underscoring the necessity to refine the summaries.

### 2.1.2. Abstractive Text Summarization

Abstractive text summarization aims to generate a summary that contains words and phrases that may not necessarily be present in the source document. It involves reinterpreting the ideas or concepts from the original text and presenting them in a condensed and rephrased form, capturing the essential information while using different wording and structures. Transformer-based models have significantly advanced the field of abstractive text summarization and offer impressive capabilities when generating concise and coherent summaries [14]. A notable transformer-based model for abstractive text summarization is the text-to-text transfer transformer (T5) [15] model, which involves encoding the input text, including the extracted sentences, into a dense vector representation that captures the semantic meaning of the text; then, the decoder part of the T5 model generates a summary by decoding this representation into coherent and concise text. Unlike extractive methods, abstractive approaches deeply analyze the input document to comprehend its main concepts, allowing them to generate new sentences that encapsulate the document's core content.

### 2.1.3. Hybrid Text Summarization

Hybrid text summarization combines both abstractive and extractive methods, with the aim of combining the strengths of each [16]. Initially, the extractive phase selects key sentences from the input text that are deemed essential to convey its overall meaning. Following this, the abstractive phase uses these extracted sentences to generate a concise and coherent summary using abstractive techniques.

### 2.1.4. Existing Approaches for Clinical Text Summarization

Electronic health records (EHRs) serve as a comprehensive and ongoing record of a patient's health information. EHR daily progress notes are written by the healthcare providers in order to track the ongoing progress of the patient on a daily basis; they serve as a chronological record of the patient's health status and play a crucial role in facilitating communication among healthcare providers and monitoring the patient's progress throughout their healthcare journey [17].

EHR daily progress notes typically follow the SOAP structure, which stands for subjective, objective, assessment, and plan. It is a documentation method designed by Larry Weed [18] to present patient's problems in a highly structured way. Within each section, there are multiple components that focus on different aspects of the patient's case and present relevant information. The *Subjective* section records subjective information obtained from the patient or caregiver, such as chief complaints, symptoms, and relevant medical history. The *Objective* section contains measurable data like vital signs, examination findings, and test results. The *Assessment* section summarizes the healthcare provider's diagnosis, evaluation of symptoms, and overall clinical impressions. Finally, the *Plan* outlines the proposed treatment plan, including medications, procedures, referrals, follow-up appointments, and patient education [19].

Recent research on EHR text summarization can be classified by the approach used. For the extractive summarization techniques, Liang et al. [20] propose a disease-specific summarization task that extracts sentences from progress notes, focusing primarily on progress notes by physicians and nurse practitioners; the pipeline includes a basic natural language processing (NLP) layer along with additional EHR-specific components such as note section classification, disease context identification, and adverse drug event detection.

HARVEST [21], an EHR summarizer deployed in a New York hospital, incorporates a Markov chain named-entity tagger to identify diseases that are explicitly mentioned in clinical notes as well as a TF-IDF scorer, which assesses the importance of these disease mentions by considering their frequency in the document and their rarity across the entire dataset.

For the abstractive summarization techniques, several works focus on summarizing radiology reports into an impression, which is a short piece of text that states the findings from the source image [22,23]. Yim and Yetisgen-Yildiz [24] present a technique for generating snippets, which are concise summaries extracted from aligned sentences. The snippets are designed to capture the essential information from the clinic visit and provide a condensed representation for easier reference and understanding.

Gao et al. [19] introduce a novel approach to automatically summarize patients' medical problems from hospital progress notes. Their experiments involved fine-tuning two pre-trained sequence-to-sequence models: T5 and BART. The study demonstrates promising results, highlighting the potential of fine-tuning the pre-trained T5 model for accurate and efficient problem summarization in the healthcare domain. It also discusses future research directions to improve the generalizability of the approach and improve the quality of generated summaries, and it proposes a manually annotated dataset based on a subset of the large and publicly available MIMIC-III database [25].

Based on these findings, the BioNLP Workshop 2023 launched a shared task on problem list summarization in January 2023 [26]. This task aimed to foster research on building NLP models for real-world diagnostic decision support in order to enhance healthcare providers' decision-making and patient care quality. Participants were tasked with developing models to generate lists of diagnoses and problems from the EHR daily progress notes of critically ill patients. Eight teams submitted their systems for evaluation, including us. The best-performing system, CUED [27], achieved an F-score of 32.77% by using an ensembled clinical T5 model.

Our work is based on the findings of Gao et al. [19]. We attempt to leverage the baseline results by experimenting with both extractive and abstractive summarization methods in a hybrid technique. Additionally, we explore the use of structured features within the *Objective* section from the same dataset. Ultimately, the main goal is to predict the ICU patients' LOSs.

## 2.2. ICU Length of Stay Prediction

ICUs play a critical role in the healthcare system by providing specialized care to patients with severe or life-threatening conditions. One key aspect of managing ICU resources and ensuring optimal patient care is predicting the LOSs in the ICU. Accurate prediction of discharge dates improves bed hour estimation, resulting in higher average occupancy and reduced waste of hospital resources [28].

Prior studies have investigated LOS prediction in disease-specific or population-specific groups such as patients with heart failure [29] or thermal burns [30] and cardiac surgery patients [31], as well as those admitted to neonatal care units [32] or ICUs [33]. While previous work has aimed to group patients based on their medical conditions, assuming each has a predefined, recommended LOS, the LOS is influenced by multiple factors beyond medical conditions. These factors include patient characteristics, presenting complaints, complications, discharge planning, and treatment complexity, all of which can extend the original target LOS. Hence, a model capable of reliably predicting patient LOS during a single visit event could help healthcare services implement preventive measures to avoid LOS extension [34], especially in a critical and costly environment such as the ICU. Few studies have used the MIMIC-III dataset for classifying ICU patient LOSs.

Wang et al. [35] introduce an open-source pipeline, MIMIC-Extract, that facilitates cohort selection and pre-processing in clinical prediction tasks using the MIMIC-III dataset. The pipeline generates a versatile cohort with diverse demographic and admission coverage, enabling various prediction tasks. It employs fixed input windows and dynamic targets to

prevent temporal label leakage in classification tasks such as mortality and length-of-stay prediction. Benchmark tasks are profiled using different machine learning (ML) models and demonstrate high performance, with the best-performing models reaching an accuracy of 69.5% for “LOS > 3 days”.

Pellegrini et al. [36] present an innovative method for patient-level predictions using graph-based unsupervised pre-training, where patients are represented as nodes in a graph, with edges indicating similarities in clinical features and histories. The model undergoes unsupervised pre-training on the patient population graph, enhancing its generalization capabilities without labeled outcomes. Fine-tuning for specific prediction tasks such as patient mortality or disease progression and LOS further improves the accuracy and robustness, and an evaluation on clinical datasets demonstrated significant enhancements in prediction accuracy over baseline models, with an accuracy of 71.44% for the “LOS > 3 days” classification task on the MIMIC-III dataset.

Classification models predict the probability of a patient staying in the ICU for a specific duration, often categorized as a “short stay” if less than 3 days and a “long stay” otherwise, and the widely used metrics for evaluating the performance of a classification model are accuracy, precision, recall, F1-score, and the area under the receiver operating characteristics curve (AUROC) [37].

For classification tasks, three prominent ML methods stand out. The MLPClassifier, a neural-network-based classification algorithm, consists of interconnected layers of neurons that apply non-linear activation functions to model complex relationships between features and target variables. Trained through back propagation, it adjusts weights to minimize the difference between predicted and actual outputs, making it popular for classification tasks [38].

Support vector machines (SVMs) aim to find an optimal hyperplane that maximizes the margin between data points of different classes and handle both linear and non-linear problems through various kernel functions [39].

Random forest, an ensemble method, combines multiple decision trees to achieve accurate classification by independently predicting class labels and aggregating results through majority voting or averaging. Known for its robustness and performance, it handles high-dimensional datasets effectively, making it less prone to overfitting while accommodating missing values and outliers [40].

This work aims to leverage the accuracy of ML classification models for the “LOS > 3 days” classification task by using the main problems and diagnoses of patients admitted to the ICU along with other structured clinical and demographic data, such as the patients’ genders, insurance details, discharge location, etc.

### 3. Materials and Methods

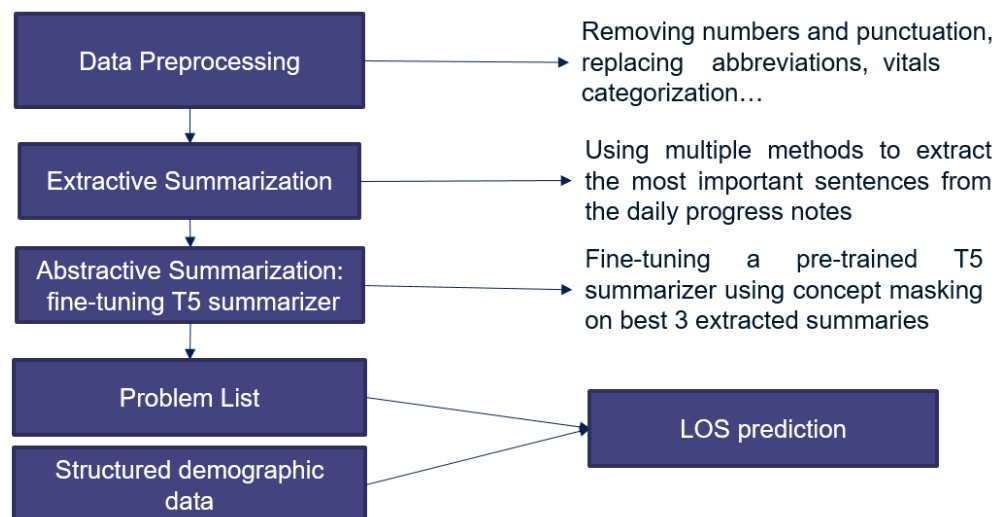
This section outlines the architecture of the proposed method and presents the materials and individual tools and techniques used in this work, including the used dataset. The overall experiment setup and exact experiments performed are presented in Section 4.

#### 3.1. The Proposed Approach

The aim of this study is to predict ICU patients’ LOSs using as input their main diagnoses and problems along with structured demographic and clinical data such as age, gender, vital signs, etc. These diagnoses are extracted from the patients’ EHR daily progress notes through summarization.

In pursuit of this objective, and as shown in Figure 1, our approach begins with summarizing EHR progress notes using NLP methods, focusing on extracting main problems and diagnoses; this method was submitted to the “BioNLP Workshop 2023: Problem List Summarization” challenge [10]. This task involves identifying and generating patient problems and diagnoses from the progress notes that are taken daily during the patient’s stay in the ICU by doctors, nurses, etc, and is known as problem list summarization.





**Figure 1.** Architecture of the proposed method.

The proposed summarization method consists of a hybrid technique wherein the extractive summarization step aims to extract the most important sentences from the EHR daily progress notes using several tools and techniques that will be introduced in this section, and a pre-trained T5 model that has been fine-tuned using concept masking performs the abstractive summarization.

Following completion of the summary of the problem list, we proceed with the extraction of additional features from the dataset: the demographic data and the LOSs of the ICU patients are mapped to the EHR daily progress notes dataset. These features, such as gender, discharge location, and other relevant factors, contribute to a comprehensive understanding of the medical profile of each patient.

Finally, the generated diagnoses and the extracted variables are used as features to predict the patients' LOSs in the ICU. Our method represents a novel approach by integrating vital signs and laboratory results into the first hybrid summarization technique proposed in this context and aimed at predicting ICU LOSs.

### 3.2. The Dataset

The dataset we used was sourced from MIMIC-III, which is a publicly available database of de-identified EHR data from approximately 60,000 hospital ICU admissions at Beth Israel Deaconess Medical Center in Boston, MA, USA, collected over an 11-year period from 2001 to 2012 [25]. This database consists of several tables containing a wide range of patient care data. These tables include *Patients*, which stores demographic information such as gender, age, and admission dates; *Admissions*, which provides details about patient admissions to the ICU, including admission and discharge diagnoses, admission types, and insurance information; and *NoteEvents*, which contains textual documentation including progress notes and discharge summaries. The training set used in this work is a sampled subset of 768 progress notes and manually annotated text spans for the SOAP components [18] from the *NoteEvents* table. The goal of the annotation was to obtain lists of problems from the *Plan* subsection. The reference summary is a list of problems mentioned in each *Plan* section that are relevant to the reasons for hospitalization. The testing set that was used to test all of the models in this work is an additional 237 annotated daily progress notes sourced from the MIMIC-III database.

### 3.3. QuickUMLS

QuickUMLS [41] is a state-of-the-art open-source system for extracting medical concepts from clinical texts and leverages the Unified Medical Language System (UMLS) through a combination of rule-based and machine learning methods. Developed by the

National Library of Medicine, UMLS provides a vast repository of medical terminology and knowledge, which QuickUMLS utilizes for accurate concept recognition. When given a text, it returns a list of UMLS concepts found in the text, including their similarities to a query string and/or other associated information.

### 3.4. TF-IDF

TF-IDF is a widely used technique in information retrieval and text mining for extractive summarization. It assigns weights to words in a document based on their frequency in the document and their rarity in the corpus. The TF-IDF score for a word is calculated, as shown in Equation (1), by multiplying its term frequency (TF) in the document by the inverse document frequency (IDF) across the corpus [42]. The TF represents the frequency of a term in a document and measures the importance of a term within a document, and the IDF measures the rarity of a term across the corpus.

$$TF - IDF = TF * IDF \quad (1)$$

### 3.5. Cosine Similarity

Cosine similarity is a metric that is commonly used in text mining, natural language processing, and information retrieval and quantifies the similarity between two vectors in a multi-dimensional space. By measuring the cosine of the angle between the vectors, it produces a value ranging from 0 to 1, where a value closer to 1 signifies higher similarity, whereas values closer to 0 indicate dissimilarity. As shown in Equation (2), the calculation of cosine similarity  $\text{sim}(a, b)$  between two vectors  $a$  and  $b$  is the dot product of the two vectors divided by the product of their magnitudes [43].

$$\text{sim}(\vec{a}, \vec{b}) = \frac{\vec{a} * \vec{b}}{\|\vec{a}\| \|\vec{b}\|} \quad (2)$$

### 3.6. Page Rank

Originally designed for ranking web pages, PageRank can also effectively rank sentences or paragraphs in extractive summarization tasks [44]. PageRank uses a sentence graph to evaluate the significance of sentences for summarization, where each sentence serves as a node, with their connections determining their importance. Edges between sentences are obtained based on criteria like semantic similarity, co-occurrence, or cosine similarity. Sentences with higher PageRank scores, indicating greater influence within the graph, are prioritized during summary generation and contribute significantly to the overall content.

### 3.7. T5 Fine-Tuning

Pre-trained models like T5 serve as a foundation for various NLP tasks but require fine-tuning to adapt them to specific domains. Fine-tuning involves training the pre-trained T5 model on task-specific data to enhance its performance for the target task [15].

To begin the fine-tuning process, the pre-trained model, which has learned general language patterns from a large text corpus, undergoes parameter adjustment using a smaller dataset specific to the task. This dataset contains input examples and corresponding outputs relevant to the target task, such as text classification, named entity recognition, or machine translation.

During fine-tuning, the pre-trained model is exposed to the task-specific dataset, and its parameters are updated using optimization techniques like gradient descent. The goal is to minimize a task-specific loss function that measures the discrepancy between the model's predictions and the true outputs in the dataset. Through fine-tuning, the model learns task-specific patterns and relationships to make accurate predictions or generate appropriate outputs.

Concept masking is a technique employed during T5 fine-tuning and is particularly useful for tasks involving generating outputs based on specific concepts or entities. Instead of traditional token replacement, concept tokens (e.g., “CONCEPT-1”, “CONCEPT-2”, etc.) are used as placeholders for desired concepts or entities. During fine-tuning, these concept tokens are replaced with the corresponding concepts from the training data, enabling the model to learn accurate concept generation based on the given inputs.

Concept masking empowers the T5 model to be trained to generate outputs conditioned on specific concepts or entities. This technique proves beneficial for tasks like text summarization, question answering, and text generation where precise control over the generated content is essential. By incorporating concept masking during fine-tuning, the model can learn to generate accurate outputs based on the provided concepts or entities.

### 3.8. Text Summarization Evaluation Metrics

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a metric set that is used to assess the quality of text summarization systems’ outputs. It counts overlapping units like n-grams and word sequences between the generated summary and a reference summary. ROUGE-L, a ROUGE variant, employs the longest common subsequence (LCS), which is the longest word sequence appearing in both summaries, regardless of consecutiveness but keeping the order. Based on LCS, Equations (3)–(5) define  $F_{LCS}$ , the LCS-based F-measure, which can be used to estimate the similarity between two translations,  $X$  of length  $m$  and  $Y$  of length  $n$ , where  $\beta = P_{LCS}/R_{LCS}$  [45].

$$R_{LCS} = \frac{LCS(X, Y)}{m} \quad (3)$$

$$P_{LCS} = \frac{LCS(X, Y)}{n} \quad (4)$$

$$F_{LCS} = \frac{(1 + \beta^2)R_{LCS}P_{LCS}}{R_{LCS} + \beta^2P_{LCS}} \quad (5)$$

## 4. Experiments and Results

This section discusses the experiments conducted as part of this work, along with their corresponding results.

### 4.1. Data Preparation

In this work, the three sections of the progress notes *Subjective*, *Objective* and *Assessment* were included. To prepare the raw dataset for analysis, several pre-processing techniques were applied in this step. Lowercasing ensured consistent casing throughout the document. Censored pattern removal addressed personally identifiable information by replacing identifiers with empty spaces to avoid interference with processing tasks. Time format removal eliminated irrelevant expressions like “HH:MM AM/PM” in order to focus on essential information. Float number representation standardization ensured uniformity by standardizing numeric formats. Text replacements improved expressions like age, medical history, and gender to enhance readability. Line break and punctuation removal maintained text coherence and reduced noise, while stemming with the Lancaster stemmer normalized words to their base form to aid with vocabulary reduction and improve text processing efficiency.

Vital signs are essential indicators of a patient’s physiological status and play a key role in diagnosing medical conditions. Unlike previous systems, this work includes vital sign measurements such as “TCO2”, which measures dissolved carbon dioxide in the blood, “glucose”, which indicates blood sugar levels, and “PLT” for assessing platelet count, among others. Each vital sign was extracted from the *Objective* section of the progress notes and was categorized based on specific thresholds according to multiple studies [46–48] to aid with interpretation and decision-making. Each category represents a range of values that



are considered “normal”, “low”, or “high” or other specific conditions depending on the laboratory result.

#### 4.2. Extractive Summarization

In the extractive summarization phase, six methods were employed to summarize the daily progress notes, including two methods that we proposed based on the combination of existing approaches.

The concept-based approach involved identifying relevant medical concepts using the QuickUMLS library and constructing a similarity matrix on concept overlap between sentences to create a graph representation with sentences as nodes and edges reflecting concept similarity and selecting the seven top-ranked sentences to form the summary using PageRank. On the other hand, the graph-based methods utilized TF-IDF vectorization and cosine similarity to construct a graph representation of the sentences, and PageRank was used to select the top seven sentences. Based on these two methods, we developed a third one whereby the graph is formed from the cosine similarity matrix of the sentences, the nodes are sentences, and the edges are the cosine similarity of the sentences based on the concepts' similarity scores; these concepts were extracted using QuickUMLS.

In the topic-based technique, sentences were tokenized and transformed into numerical vectors using TF-IDF values, and cosine similarity was employed to select the seven top-ranked sentences. We proposed another technique that combined concept extraction with TF-IDF, whereby concepts were extracted using QuickUMLS, and TF-IDF scores were calculated alongside concept frequencies to determine sentence importance; the model ultimately selected the seven top-ranked sentences.

Lastly, the clustering-based method employed BERT, which is a transformer-based model: specifically, the “distilbert-base-uncased” variant, with the “hidden” parameter set to “[−1, −2]”. The number of sentences to be present in the summary was set to seven.

These models were trained on a “Google Colab” notebook for approximately 3 h. We then submitted them for testing to a CodaLab competition (CodaLab Competition for the BioNLP Workshop 2023 Shared Task 1A: Problem List Summarization, 2023, <https://codalab.lisn.upsaclay.fr/competitions/12388> (accessed 27 June 2024)) that was created for the “BioNLP Workshop 2023 Shared Task 1A: Problem List Summarization” challenge in order to manage system submissions and show the leaderboard. The CodaLab was configured to receive system prediction outputs in a text file format, and it ran a ROUGE-L evaluation script to produce scores. Our results are summarized in Table 1. The concept-based approach achieved a relatively high precision of 1.35%, indicating that the selected summary sentences closely matched the reference summaries. However, its recall of 41.87% suggests that it may have missed some important information, resulting in a lower F-score of 2.56%. The graph-based method, on the other hand, performed poorly in terms of precision (0.0006%) and F-score (0.001%). This suggests that the selected summary sentences may not be relevant to the reference summaries. Although it exhibited a higher recall of 0.08%, it failed to produce meaningful summaries overall. Combining the concept and graph approaches in the concept + graph-based method showed a slight improvement in precision (1.36%) and F-score (2.59%) compared to concept-based alone. However, its recall remained similar to concept-based, indicating that the addition of graph-based did not significantly enhance the quality of the summaries.

**Table 1.** ROUGE scores in percentage (%) for proposed extractive summarization methods.

Method	$P_{LCS}$	$R_{LCS}$	$F_{LCS}$
Concept-based	1.35	41.87	2.56
Graph-based	0.0006	0.08	0.001
Concept + Graph-based	1.36	41.74	2.59
Topic-based	1.29	42.89	2.47
Topic + Concept-based	0.64	10.73	1.19
Clustering-based	1.69	38.91	3.17

The topic-based method achieved a relatively high recall of 42.89%, suggesting that it successfully captured most of the important information from the reference summaries. However, its precision of 1.29% and F-score of 2.47% were comparatively lower, indicating that it may have generated some irrelevant sentences in the summary. When topic and concept approaches were combined in the topic + concept-based method, precision decreased significantly to 0.64%, indicating a decrease in the quality of selected summary sentences. The recall (10.73%) and F-score (1.19%) also dropped, indicating that the combination did not effectively improve the performance compared to topic-based alone.

Lastly, the clustering-based method achieved the highest precision (1.69%) and F-score (3.17%) among the evaluated methods. This indicates that it selected summary sentences that closely matched the reference summaries, resulting in higher-quality summaries. However, its recall (38.91%) was relatively low, suggesting it may have missed some relevant information from the source document.

It is observed that the overall performance of these models is very low; this is due to the fact that the summaries generated by these models were tested against the ground truth as annotated in [19]. The diagnoses and problems presented in the ground truth do not appear explicitly in the notes because they were manually extracted [19]. For this reason, the ROUGE scores, which test the similarity of the generated summaries and the ground truth by counting the overlapping words and concepts in both sequences, were very low for the extractive summarization models. This highlights the need for the abstractive summarization step. The goal of this step was to shorten the text and keep only the most important sentences.

#### 4.3. Abstractive Summarization

For this step, we used the outputs of the three best extractive summarization methods, based on their ROUGE-L F-scores, as input to the T5-base model. We worked on fine-tuning the pre-trained model using the concept masking technique. The hyperparameters of the T5-base model included an “AdamW” optimizer and a learning rate of  $1 \times 10^{-4}$ . The number of training epochs was only two, and the “max input size” was set to “512”. We ran the models locally for two epochs on a NVIDIA GeForce GTX Ti GPU with Max-Q Design for a total of about 8 h. Similarly to the extractive summarization step, we submitted the models for testing to the CodaLab competition, with our best-performing hybrid summarizer displayed on the leaderboard: (CodaLab leaderboard of the BioNLP Workshop 2023 Shared Task 1A: Problem List Summarization, 2023, <https://codalab.lisn.upsaclay.fr/competitions/12388#results> (accessed 27 June 2024)). The results are shown in Table 2.

**Table 2.** ROUGE scores in percentage (%) for proposed hybrid summarization methods in comparison with baseline and benchmark methods.

Method	$P_{LCS}$	$R_{LCS}$	$F_{LCS}$
Baseline [19]	-	-	15
Graph + Concept + T5	39.44	15.18	19.66
BERT + T5	40.19	15.41	20.30
Concept + T5	43.83	14.99	20.32
CUED [27]	41.69	30.51	32.77

All of the methods performed better than the baseline [19], with the combination of concept-based summarization and T5 summarization being our best proposed method: achieving a precision of 43.83%, a recall of 14.99%, and an F-score of 20.32%. Its generated summaries are to be used as features for LOS prediction. While the benchmark method [27] is still at the top of existing methods with an F-score of 32.77%, our best method surpasses its precision of 41.69%.

It was observed that the precision values were higher than the recall values. This suggests that the model was better at avoiding false positives, which means that it accurately identified instances that were not relevant. However, it may have missed a significant number of true positive instances, indicating that some relevant information was not captured.

There are two possible reasons for the low recall. Firstly, the abstractive method might have failed to generate the necessary concepts. This is logical considering the small dataset and the absence of any data augmentation techniques, which could have limited the model's ability to generate comprehensive summaries. Secondly, the extractive method might have removed sentences containing important information required for the abstractive summarization step. This implies that during the extraction process, some relevant sentences were not retained, resulting in a loss of vital information in the generated summaries.

The trade-off between precision and recall depends on the specific requirements of the summaries. In this case, as the summaries are used for ICU LOS prediction, having high precision is desirable. This means that the generated summaries are more likely to provide accurate and relevant information, enabling more precise predictions. Therefore we deem the output from the concept-based and T5 summarization combination method, which has higher precision than the benchmark method, to be suitable as input for the next step of the ICU LOS classification.

#### 4.4. Length of Stay Classification

In this study, we focused on classifying ICU length of stay (LOS) into two classes: "short" if the LOS is 3 days or less and "long" otherwise. We then compared our proposed systems with existing methods [35,36] for the same classification task. Wang et al. [35] achieved an accuracy of 69.50%, an F-score of 59.50%, and an AUROC of 73.60%. In comparison, Pellegrini et al. [36] achieved a higher accuracy of 71.44% and the highest AUROC among the benchmark methods at 77.78%. These results indicate that, while both benchmark methods are strong, Pellegrini et al.'s model is particularly robust in terms of overall predictive performance, as indicated by the AUROC metric.

The features we used included problem list summaries, which were the main diagnoses relevant to the reasons for hospitalization that were present but not explicit in the *Plan* section of the EHR daily progress notes; these were extracted manually for the challenge [19]. We used diagnoses we extracted as well from the other sections of the EHR daily progress notes and other clinical demographic features from the MIMIC-III database. These features were mapped across other tables such as *Patients*, *ICUStays*, and *Admissions* using the "HADM\_ID" column, which stands for each hospital admission of a patient. Included features were LOS, our target, insurance information, gender, ethnicity, religion, admission type, first and last care unit before ICU, and discharge location.

The features were encoded and fed into three classifiers: SVMClassifier, RandomForestClassifier (RF), and MLPClassifier (MLP) from the scikit-learn library [49]. Diagnoses and problems were encoded using QuickUMLS to extract medical concepts from the dataset. The classifiers were trained using five-fold cross-validation for hyperparameter tuning and were tested on the challenge's testing dataset. Four experiments were run using different combinations of features. The results are shown in Table 3.

When using only clinical and demographic features, our proposed models showed varying degrees of performance. The MLP model achieved an accuracy of 63.00%, a precision of 65.13%, a recall of 82.50%, an F-score of 72.79%, and an AUROC of 58.60%. The RF model had a similar accuracy of 63.00% and a slightly higher precision of 65.33% and a recall of 81.67%, resulting in an F-score of 72.59% and an AUROC of 62.50%. The SVM model outperformed the other models in this category, achieving the highest accuracy of 64.50%, the highest precision of 67.13%, a recall of 80.00%, an F-score of 73.00%, and the highest AUROC of 63.30%. These results indicate that while SVM shows the best performance among models using only clinical and demographic features, it still falls short compared to the benchmark methods in terms of AUROC.

**Table 3.** Performance comparison of proposed systems and benchmark methods in term of accuracy, precision, recall, F1, and AUROC scores in percentages (%).

	Method	Accuracy	Precision	Recall	F-Score	AUROC
Benchmark Methods	Wang et al. [35]	69.50	-	-	59.50	73.60
	Pellegrini et al. [36]	71.44	-	-	-	77.78
Clinical and Demographic features only	MLP	63.00	65.13	82.50	72.79	58.6
	RF	63.00	65.33	81.67	72.59	62.50
	SVM	64.50	67.13	80.00	73.00	63.3
Summaries only	MLP	62.50	66.42	75.83	70.81	67.50
	RF	58.50	60.11	91.67	72.61	63.80
	SVM	66.00	68.57	80.00	73.84	69.30
Diagnoses only	MLP	70.00	74.59	75.83	75.21	76.80
	RF	69.50	69.28	88.33	77.66	70.20
	SVM	72.50	79.27	73.33	76.19	76.20
Diagnoses and summaries	MLP	72.50	74.07	83.33	78.43	79.70
	RF	66.00	64.13	98.33	77.63	81.70
	SVM	77.50	78.19	86.66	82.21	82.40

When using only summaries as features, the performance of the models varied significantly. The MLP model achieved an accuracy of 62.50% and a precision of 66.42%, a recall of 75.83%, an F-score of 70.81%, and an AUROC of 67.50%. The RF model had a lower accuracy at 58.50% but achieved the highest recall of 91.67% among all models using summaries, resulting in an F-score of 72.61% and an AUROC of 63.80%. The SVM model again performed the best in this category, achieving the highest accuracy of 66.00%, the highest precision of 68.57%, a recall of 80.00%, the highest F-score of 73.84%, and the highest AUROC of 69.30%. These results suggest that summaries alone can be highly predictive, with the SVM model particularly excelling using this feature set.

The models using only diagnoses as features performed notably well. The MLP model achieved an accuracy of 70.00% and a precision of 74.59%, a recall of 75.83%, an F-score of 75.21%, and the highest AUROC in this category of 76.80%. The RF model had a similar accuracy of 69.50% and a precision of 69.28%, the highest recall of 88.33% among all models using diagnoses, an F-score of 77.66%, and an AUROC of 70.20%. The SVM model outperformed both, achieving the highest accuracy of 72.50%, the highest precision of 79.27%, a recall of 73.33%, an F-score of 76.19%, and an AUROC of 76.20%. These results demonstrate the strong predictive performance of diagnosis features, with the SVM model again showing superior performance.

Combining diagnoses and summaries yielded the best results among all of our proposed models. The MLP model achieved an accuracy of 72.50%, a precision of 74.07%, a recall of 83.33%, an F-score of 78.43%, and an AUROC of 79.70%. The RF model had a lower accuracy at 66.00% and a precision of 64.13%, the highest recall of 98.33% among all models, an F-score of 77.63%, and an AUROC of 81.70%. The SVM model performed the best overall, achieving the highest accuracy of 77.50%, the highest precision of 78.19%, a recall of 86.66%, the highest F-score of 82.21%, and the highest AUROC of 82.40%. These results indicate that combining diagnoses and summaries provides the most comprehensive and accurate predictive performance, with the SVM model being the most effective.

In summary, our proposed methods show varying degrees of success, with models using diagnoses and summaries combined outperforming models that use other feature sets. The SVM model consistently performed the best across different feature sets, especially when both diagnoses and summaries were used; the best hyperparameters found for this model were “C = 1” for the regularization parameter, “gamma = 0.01” for the kernel coefficient, and a “radial basis function (RBF)” as the kernel. This model, being our best-performing classifier, has set a new benchmark for this task and showed significant improvements and robustness over the benchmark methods (*Papers With Code*, Length of Stay Prediction on MIMIC-III, 2022, <https://paperswithcode.com/sota/length-of-stay-prediction-on-mimic-iii> (accessed 27 June 2024)), especially in terms of precision and recall.

## 5. Conclusions and Future Work

In this work, we used NLP methods to summarize EHR progress notes from ICU patients and to extract diagnoses and problems, with the ultimate goal of predicting ICU patients' lengths of stay. Our study showcased the effectiveness of combining extractive and abstractive summarization techniques. Moreover, the generated summaries and extracted diagnoses proved valuable for predicting ICU LOS, offering potential benefits for ICU management, clinical decision-making, and patient care in real-world clinical settings.

Moving forward, exploring the integration of additional clinical variables and refining the summarization techniques could enhance the predictive accuracy and utility of EHR-driven LOS predictions. While our LOS classification methods show significant improvements and robustness over the benchmark methods, especially in terms of precision and recall, they still need further enhancement. Future work could focus on refining these models, potentially through additional feature engineering and model optimization. This could include developing temporal features that capture trends over time, such as the progression of vital signs or lab results instead of categorizing them to "low", "normal", and "high" as we did in this work. Furthermore, model optimization techniques, like implementing feature selection methods, including recursive feature elimination (RFE), can help identify the most important features from the dataset and enhance the models' predictive performance by reducing the dataset's dimensionality. Additionally, gathering feedback from healthcare professionals regarding the generated summaries and LOS predictions can provide valuable insights into their usefulness, readability, and accuracy.

## 6. Limitations

The primary challenge in generating problem list summaries from this dataset was that certain concepts present in the reference summary were not found within the other sections of the progress notes. This occurred mainly because the dataset was manually annotated for the challenge. Due to this limitation, the summarization process faced a hindrance, as important medical concepts that are necessary for extracting the main diagnoses of patients and thus generating accurate summaries were missing. As a result, the generated summaries ended up being incomplete or inaccurate due to a lack of relevant information. Furthermore, the relatively small number of data points available for training and evaluation further constrained our analysis. This limitation emphasizes the need for larger and more diverse datasets to enhance the robustness and generalization of our models. Conducting experiments with data augmentation techniques or annotating another EHR daily progress note dataset can address this issue and augment the diversity and quantity of training data, potentially improving the models' performance and generalization.

**Author Contributions:** Conceptualization, S.R., Y.D. and N.S.N.; methodology, S.R., Y.D. and N.S.N.; software, S.R.; validation, S.R., F.C., S.C.-D., Y.D., S.K. and N.S.N.; formal analysis, S.R.; investigation, S.R., Y.D. and N.S.N.; writing—original draft preparation, S.R.; writing—review and editing, S.R., F.C., S.C.-D., Y.D., S.K. and N.S.N.; visualization, S.R.; supervision, S.C.-D., Y.D. and N.S.N.; project administration, S.R. and N.S.N.; funding acquisition, S.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was conducted with the financial support of Erasmus+ ICM, funded by the European Union, project number 2020-1-IE02-KA107-000730, and the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under grant No. 18/CRT/6223. The views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. The study used the MIMIC-III clinical dataset, which is available at <https://physionet.org/>, and an annotated subset



of the database that was introduced at the BioNLP Workshop 2023: Problem List Summarization, which is available at <https://physionet.org/content/bionlp-workshop-2023-task-1a/2.0.0/>.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

AUROC	area under the receiver operating characteristics curve
BERT	bidirectional-encoder transformer
EHR	electronic health records
ICU	intensive care unit
LCS	longest common subsequence
LOS	length of stay
MIMIC-III	Medical Information Mart for Intensive Care III
ML	machine learning
MLP	multilayer perceptron
NLP	natural language processing
RBF	radial basis function
RFE	recursive feature elimination
RF	random forest
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SOAP	subjective objective assessment plan
SVM	support vector machine
T5	text-to-text transfer transformer
TF-IDF	term frequency–inverse document frequency
UMLS	Unified Medical Language System

## References

1. Stone, K.; Zwiggelaar, R.; Jones, P.; Mac Parthaláin, N. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLoS Digital Health* **2022**, *1*, e0000017. [[CrossRef](#)]
2. OECD. *Health at a Glance 2023*; OECD Indicators, OECD Publishing: Paris, France, 2023; p. 234. [[CrossRef](#)]
3. Ohsfeldt, R.L.; Choong, C.K.C.; Mc Collam, P.L.; Abedtash, H.; Kelton, K.A.; Burge, R. Inpatient hospital costs for COVID-19 patients in the United States. *Adv. Ther.* **2021**, *38*, 5557–5595. [[CrossRef](#)] [[PubMed](#)]
4. Iwase, S.; Nakada, T.; Shimada, T.; Oami, T.; Shimazui, T.; Takahashi, N.; Yamabe, J.; Yamao, Y.; Kawakami, E. Prediction algorithm for ICU mortality and length of stay using machine learning. *Sci. Rep.* **2022**, *12*, 12912. [[CrossRef](#)] [[PubMed](#)]
5. Teno, J.M.; Fisher, E.; Hamel, M.B.; Wu, A.W.; Murphy, D.J.; Wenger, N.S.; Lynn, J.; Harrell, F.E., Jr. Decision-making and outcomes of prolonged ICU stays in seriously ill patients. *J. Am. Geriatr. Soc.* **2000**, *48*, S70–S74. [[CrossRef](#)] [[PubMed](#)]
6. Toh, H.; Lim, Z.; Yap, P.; Tang, T. Factors associated with prolonged length of stay in older patients. *Singap. Med. J.* **2017**, *58*, 134–138. [[CrossRef](#)] [[PubMed](#)]
7. Inabnit, L.S.; Blanchette, C.; Ruban, C. Comorbidities and length of stay in chronic obstructive pulmonary disease patients. *COPD J. Chronic Obstr. Pulm. Dis.* **2018**, *15*, 355–360. [[CrossRef](#)] [[PubMed](#)]
8. Furlow, B. Information overload and unsustainable workloads in the era of electronic health records. *Lancet Respir. Med.* **2020**, *8*, 243–244. [[CrossRef](#)] [[PubMed](#)]
9. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [[CrossRef](#)] [[PubMed](#)]
10. Gao, Y.; Dligach, D.; Miller, T.; Afshar, M. BioNLP Workshop 2023 Shared Task 1A: Problem List Summarization. In Proceedings of the 22nd Workshop on Biomedical Language Processing, Toronto, ON, Canada, 13 July 2023. [[CrossRef](#)]
11. El-Kassas, W.S.; Salama, C.R.; Rafea, A.A.; Mohamed, H.K. Automatic text summarization: A comprehensive survey. *Expert Syst. Appl.* **2021**, *165*, 113679. [[CrossRef](#)]
12. Moratanch, N.; Chitrakala, S. A survey on extractive text summarization. In Proceedings of the 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 10–11 January 2017; pp. 1–6. [[CrossRef](#)]
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:cs.CL/1810.04805.
14. Ranganathan, J.; Abuka, G. Text Summarization using Transformer Model. In Proceedings of the 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), Milan, Italy, 29 November–1 December 2022; pp. 1–5. [[CrossRef](#)]

15. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2023**, arXiv:cs.LG/1910.10683.
16. Daga, G.; Saha, S.; Shah, Y.; Nirmala, S.J. Abstractive Text Summarization Using Hybrid Methods. In Proceedings of the 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 11–12 August 2022; pp. 1294–1300. [\[CrossRef\]](#)
17. Shoolin, J.; Ozeran, L.; Hamann, C.; Bria Li, W. Association of Medical Directors of Information Systems consensus on inpatient electronic health record documentation. *Appl. Clin. Inform.* **2013**, *4*, 293–303.
18. Weed, L.L. Medical records, patient care, and medical education. *Ir. J. Med. Sci. (1926–1967)* **1964**, *39*, 271–282. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Gao, Y.; Dligach, D.; Miller, T.; Xu, D.; Churpek, M.M.M.; Afshar, M. Summarizing Patients' Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 2979–2991.
20. Liang, J.; Tsou, C.H.; Poddar, A. A novel system for extractive clinical note summarization using EHR data. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 46–54.
21. Hirsch, J.S.; Tanenbaum, J.S.; Lipsky Gorman, S.; Liu, C.; Schmitz, E.; Hashorva, D.; Ervits, A.; Vawdrey, D.; Sturm, M.; Elhadad, N. HARVEST, a longitudinal patient record summarizer. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 263–274. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Zhang, Y.; Ding, D.Y.; Qian, T.; Manning, C.D.; Langlotz, C.P. Learning to summarize radiology findings. *arXiv* **2018**, arXiv:1809.04698.
23. Sotudeh Gharebagh, S.; Goharian, N.; Filice, R. Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1899–1905. [\[CrossRef\]](#)
24. Yim, W.W.; Yetisgen-Yildiz, M. Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization. In Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, Online, 6 June 2021; pp. 10–20.
25. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.W.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [\[CrossRef\]](#)
26. Gao, Y.; Dligach, D.; Miller, T.; Afshar, M. Overview of the Problem List Summarization (ProbSum) 2023 Shared Task on Summarizing Patients' Active Diagnoses and Problems from Electronic Health Record Progress Notes. In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, ON, Canada, 13 July 2023; pp. 461–467. [\[CrossRef\]](#)
27. Manakul, P.; Fathullah, Y.; Liusie, A.; Raina, V.; Raina, V.; Gales, M. CUED at ProbSum 2023: Hierarchical Ensemble of Summarization Models. In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, ON, Canada, 13 July 2023; pp. 516–523. [\[CrossRef\]](#)
28. Azari, A.; Janeja, V.P.; Mohseni, A. Predicting Hospital Length of Stay (PHLOS): A Multi-tiered Data Mining Approach. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops, Brussels, Belgium, 10 December 2012; pp. 17–24. [\[CrossRef\]](#)
29. Whellan, D.J.; Zhao, X.; Hernandez, A.F.; Liang, L.; Peterson, E.D.; Bhatt, D.L.; Heidenreich, P.A.; Schwamm, L.H.; Fonarow, G.C. Predictors of hospital length of stay in heart failure: Findings from Get With the Guidelines. *J. Card. Fail.* **2011**, *17*, 649–656. [\[CrossRef\]](#)
30. Hussain, A.; Dunn, K.W. Predicting length of stay in thermal burns: A systematic review of prognostic factors. *Burns* **2013**, *39*, 1331–1340. [\[CrossRef\]](#)
31. Almashrafi, A.; Alsabti, H.; Mukaddirov, M.; Balan, B.; Aylin, P. Factors associated with prolonged length of stay following cardiac surgery in a major referral hospital in Oman: A retrospective observational study. *BMJ Open* **2016**, *6*, e010764. [\[CrossRef\]](#)
32. Seaton, S.E.; Barker, L.; Jenkins, D.; Draper, E.S.; Abrams, K.R.; Manktelow, B.N. What factors predict length of stay in a neonatal unit: A systematic review. *BMJ Open* **2016**, *6*, e010466. [\[CrossRef\]](#)
33. Atashi, A.; Ahmadian, L.; Rahmatinezhad, Z.; Miri, M.; Nazeri, N.; Eslami, S. Development of a national core dataset for the Iranian ICU patients outcome prediction: A comprehensive approach. *BMJ Health Care Inform.* **2018**, *25*, 71–76. [\[CrossRef\]](#)
34. Gokhale, S.; Taylor, D.; Gill, J.; Hu, Y.; Zeps, N.; Lequertier, V.; Prado, L.; Teede, H.; Enticott, J. Hospital length of stay prediction tools for all hospital admissions and general medicine populations: Systematic review and meta-analysis. *Front. Med.* **2023**, *10*, 1192969. [\[CrossRef\]](#)
35. Wang, S.; McDermott, M.B.A.; Chauhan, G.; Ghassemi, M.; Hughes, M.C.; Naumann, T. MIMIC-Extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. In Proceedings of the ACM Conference on Health, Inference, and Learning, CHIL'20, Toronto, ON, Canada, 2–4 April 2020; pp. 222–235. [\[CrossRef\]](#)
36. Pellegrini, C.; Navab, N.; Kazi, A. Unsupervised pre-training of graph transformers on patient population graphs. *Med. Image Anal.* **2023**, *89*, 102895. [\[CrossRef\]](#)
37. Hossin, M.; Sulaiman, M.N. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process.* **2015**, *5*, 1.
38. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#)
39. Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.

40. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
41. Soldaini, L.; Goharian, N. Quickumls: A fast, unsupervised approach for medical concept extraction. In Proceedings of the MedIR Workshop, SIGIR, Pisa, Italy, 21 July 2016; pp. 1–4.
42. Sammut, C.; Webb, G.I. (Eds.) TF-IDF. In *Encyclopedia of Machine Learning*; Springer US: Boston, MA, USA, 2010; pp. 986–987. [[CrossRef](#)]
43. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [[CrossRef](#)]
44. Brin, S.; Page, L. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* **1998**, *30*, 107–117. [[CrossRef](#)]
45. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
46. Gunst, J.; Van den Berghe, G. Blood glucose control in the ICU: How tight? *Ann. Transl. Med.* **2017**, *5*, 76. [[CrossRef](#)]
47. Zarychanski, R.; Houston, D.S. Assessing thrombocytopenia in the intensive care unit: The past, present, and future. *Hematol. Am. Soc. Hematol. Educ. Program* **2017**, *2017*, 660–666. [[CrossRef](#)] [[PubMed](#)]
48. Kraut, J.A.; Madias, N.E. Re-Evaluation of the Normal Range of Serum Total CO<sub>2</sub> Concentration. *Clin. J. Am. Soc. Nephrol.* **2018**, *13*, 343–347. [[CrossRef](#)] [[PubMed](#)]
49. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.