electronics

MDPI

Article

Transformer-Based Prediction of Hospital Readmissions for Diabetes Patients

Jorge García-Mosquera, María Villa-Monedero, Manuel Gil-Martín and Rubén San-Segundo *

Speech Technology and Machine Learning Group, Information Processing and Telecommunications Center, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain; j.gmosquera@alumnos.upm.es (J.G.-M.); maria.villa.monedero@alumnos.upm.es (M.V.-M.); manuel.gilmartin@upm.es (M.G.-M.)

* Correspondence: ruben.sansegundo@upm.es; Tel.: +34-910672225

Abstract: Artificial intelligence is having a strong impact on healthcare services, improving their quality and efficiency. This paper proposes and evaluates a prediction system of hospital readmissions for diabetes patients. This system is based on a Transformer, a state-of-the-art deep learning architecture integrating different types of information and features in the same model. This architecture integrates several attention heads to model the contribution of each feature to the global prediction. The main target of this work is to provide a decision support tool to help manage hospital resources effectively. This system was developed and evaluated using the United States Health Facts Database, which includes information and features from 101,766 diabetes patients between 1999 and 2008. The experiments were conducted using a patient-wise cross-validation strategy, ensuring that the patients used to develop the system were not used in the final test. These experiments demonstrated the Transformer's strong ability to combine different features, providing slightly better results compared to previous results reported on this dataset. These experiments allow us to report the prediction accuracy for multiple class numbers. Finally, this paper provides a detailed analysis of the relevance of each feature when predicting hospital readmissions.

Keywords: transformer-based prediction; diabetes patients; hospital readmission prediction; feature analysis; combination of different types of features

Academic Editor: Ping-Feng Pai

Received: 17 October 2024
Revised: 12 December 2024
Accepted: 18 December 2024
Published: 3 January 2025

Citation: García-Mosquera, J.; Villa-Monedero, M.; Gil-Martín, M.; San-Segundo, R. Transformer-Based Prediction of Hospital Readmissions for Diabetes Patients. *Electronics* **2025**, *14*, 174. https://doi.org/10.3390/electronics14010174

1. Introduction

In recent decades, we have witnessed the potential and impact of artificial intelligence in improving the quality of health and sanitary systems. Some diseases, such as diabetes, are increasing in incidence, with a rise of 6% in the global population between 2000 and 2020 for adults [1]. If this trend continues, the International Diabetes Federation estimates that the prevalence will reach over 783 million people by 2045. Specifically, this growth is due to the epidemiological entrenchment of type 2 diabetes, overweight and obesity (declared by the World Health Organization), a sedentary lifestyle, and the aging of the population in some countries. Despite this growth, the development of new medications and effective therapies that preserve glycemic and cardiovascular control has facilitated a decline in mortality. Nonetheless, dealing with this emerging incidence of diabetic patients presents a management challenge because this illness involves a substantial increase in hospital admissions. Recent studies [2] reveal a relative rise in hospital admissions associated with diabetes of 6.6% in the United States between 2010 and 2019, alongside a 35.4% increase in the total cost of admissions, reaching United States Dollars 132 billion.

One of the most important aspects of the management of hospital and health resources takes place when the professionals must decide whether to discharge a previously admitted patient. This complex decision depends on multiple variables that determine the current condition of the patient and whether they fulfill the requirements for discharge, in addition to other elements such as the saturation of the hospital unit.

The preventability of the possible readmission of a patient is determined to avoid frustration and reduce costs. Furthermore, early unplanned readmissions (defined as those that take place within 30 days of discharge) are often used as a parameter of the quality of healthcare [3], especially in cases where readmissions are subsequently described as inevitable [4].

Professional decision-making tools offer a great opportunity to decrease the error rate when a patient is discharged. Recently, a variety of techniques and models have been introduced with these capacities due to the large-scale collection of anonymized data from patients. These advancements are driven by the research and development of algorithms that are able to extract the most relevant information for decision-making. Expert systems [5] are algorithms that simulate the decision-making abilities of experts, optimizing the financial and quality aspects of healthcare systems. These models offer excellent support and validation for the decision-making processes carried out by professionals, whether in diagnosis and treatment or healthcare resource management. Importantly, they also contribute significantly to improving the patient's quality of life and recovery.

The main contributions of this paper are as follows:

- The proposal and evaluation of a deep learning-based system to predict hospital readmissions of diabetes patients;
- The adaptation of a Transformer-based system, allowing the combination of different types of features or information (categorical or numerical);
- A comparison with previous work on the same dataset, obtaining better results;
- A detailed analysis of the relevance of each feature when predicting readmissions.

2. State of the Art

Risk factors that lead to hospital readmissions caused by diabetes mellitus have been studied in numerous related works. Robbins et al. [6] reviewed and identified statistically significant risk factors in diabetes. Some of the most relevant features, such as patient age or race, in addition to comorbidities like cardiovascular and renal diseases and insulin therapy [7], were included in this work. It is important to establish the clinical situations and characteristics that have the greatest impact on the readmission of diabetic patients, as their 30-day readmission rate of hospitalized patients is higher than that for the overall hospitalized population [8]. Healy et al. [9] explored more challenging factors that influence readmission rates, such as inpatient diabetes education.

Many studies develop machine learning and deep learning algorithms to predict the possible readmission of an inpatient and discuss the most relevant clinical features for classification models. Hsu et al. [10] built an integrated genetic algorithm and Support Vector Machine (SVM) to determine the readmissions of pneumonia patients within 30 days and compared the results with the performance of other models, such as logistic regression or deep neural networks. In the diabetes field, Cui et al. [11] trained and improved an SVM-based model with a genetic algorithm to achieve better results in predicting the readmission of diabetic patients. Dafrallah et al. [12] analyzed the risk factors for hospital readmission and employed machine learning, particularly gradient boosting, to predict readmissions using demographic and clinical data. As shown in these previous works, readmission decisions are complex and depend on many factors. The automatic prediction of readmission requires machine learning algorithms capable of integrating different types

of information (numerical, categorical, etc.). Conventional models [13] include feature extraction modules adapted to each type of information.

Compared to conventional models, deep learning algorithms have demonstrated better performance in classification and prediction tasks, but their performance strongly depends on the amount of available data. Integrating different sources of information in deep learning algorithms requires preprocessing modules specifically designed for each type of information. Because of this, many works focus on one type of information: for example, Li et al. [14] proposed a nonlinear aggregated graph neural network model predicting diabetes based on blood glucose data. Transformer-based models are state-of-the-art deep learning models showing very good results in a wide range of applications: Zhu et al. [15] designed a temporal fusion Transformer model, trained with personal data and embedded within a system-on-a-chip, to monitor real-time blood glucose. Vision Transformer models can also be used to analyze images in order to prevent severe complications, such as diabetic retinopathy [16]. Many of these applications only integrate one type of information in the algorithm. There are also solutions for combining structured and unstructured data (Chiu et al. [17]), for readmission prediction, but they require a hybrid model including Transformers and long short-term memory networks.

Large language models are based on Transformers and handle information expressed in words, symbols, and numbers. This paper explores the possibility of using Transformers (a state-of-the-art deep learning architecture) to integrate different types of information without a preprocessing step.

This paper uses a public dataset, for which the best results were obtained by Lu et al. [18] using stacking-based models to assemble the individual predictions of different machine learning algorithms, leading to better accuracy results. Every type of information was preprocessed by a specific module. The question addressed in this paper is whether a Transformer-based architecture can combine several types of information without specific preprocessing modules.

3. Materials and Methods

This Section describes the dataset used to train and evaluate the proposed system and the main machine learning algorithms employed.

3.1. Database

Herein, we describe the data collection and processing.

3.1.1. Data Collection

The database originates from Cerner Corporation (Kansas City, MI, USA), which provided this database through the voluntary Health Facts program, in which information was collected by Strack et al. [19]. This program gathers clinical records starting from 1999 from 130 hospitals across the USA. The repository includes data on consultations, medical specialties, diagnoses, inpatient procedures, and demographic information. All information is anonymized and contains no personally identifiable data. The database covers a period of 10 years, from 1999 to 2008. Initially, it consists of 41 non-relational tables, comprising a total of 117 attributes. In total, 74,036,643 rows of consultations corresponding to 20,769,802 patients are included.

The extraction criteria for the initial dataset used by Strack et al. are (1) pertaining to a hospitalization, (2) with a diagnosis of any type of diabetes, (3) with a length of stay between 1 and 14 days, (4) with laboratory tests conducted, and (5) medications administered during the stay. A total of 101,766 patients meet this selection, and 55 feature columns are extracted for each record as the most relevant, with the final column indicating the class as follows:

readmission within 30 days (<30), readmission after 30 days (>30), or no readmission (NO). This database was released for the research community, and it has become a benchmark for research in this area. The dataset is available online and provided by Strack et al. [19], who also include a table with the detailed definitions for each feature. This dataset was the experimental setup for the present work.

3.1.2. Data Preprocessing

Cleaning the diabetic patient database is essential to reduce the model's confusion in interpreting the data. In line with the method described in the works of Strack et al. [19] and Lu et al. [18], features with more than 30% missing data were removed: 'Weight' (97%), 'Payer Code' (40%), and 'Medical Specialty' (47%). Additionally, 'EncounterID' and 'Patient Number', which hold no medical relevance, were discarded. Columns with a single value, such as the medications 'Sitagliptin' and 'Examine', were eliminated as well, since they do not influence the model's classification. Patients with invalid race and gender variables were also excluded. Diagnostic features were integrated according to the ICD-9-CM criteria to reduce the number of values offered. Patients discharged due to death or palliative care were likewise removed, as they could not be readmitted. These decisions, made in previous studies, were replicated in this work to allow a fair comparison between the performance of previous models and the one proposed in this work.

The three labels or classes are readmission within 30 days, readmission after 30 days, and no readmission, identified as LESS, MORE, and NO, respectively. These target values will define the type of problem: for binary classification, merging readmission after 30 days and no readmission, against readmission within 30 days; for multiclass classification, the three classes are treated separately. A substantial number of studies select 30-day readmission as the threshold for early unplanned readmission [20].

Finally, the dataset was adjusted using random under-sampling to ensure that the models were fed the same number of instances for each readmission class and preventing imbalance (as there were significantly more non-readmitted patients). The lower limit was set by the LESS class, represented with 8901 entries. Accordingly, the adjusted dataset contained 17,802 records for two classes and 26,703 for three classes. Random under-sampling was also performed in this work to allow a fair comparison between the performance of previous models and the one proposed in this work.

So that the database could be input for natural language processing (NLP) as a list of sentences with a patient's information in each, the tabulated rows were transformed into text sequences with the readmission class as the last element of the sequence, in .txt format and UTF-8 encoding. Each text sequence corresponded to one row from the dataset, with the elements separated by a space and the final class separated by a tab.

3.1.3. Clinical Series

The clinical series (Table 1) is the set of clinical cases that share common characteristics, and it is of significant relevance in identifying patterns, in this case, of readmission. It helps to understand the output of the model concerning some weighty features.

Table 1. Clinical series of diagnostic, demographic, and hospitalization features.

Feature	Total Population		After Data Preprocessing	
	Percentage [1]	Readmission Rate	Percentage	Readmission Rate
Primary diagnosis				
Neoplasms	3.1%	10.7%	3.1%	48.1%
Diabetes	8.6%	12.6%	8.8%	55.0%
Circulatory	29.7%	11.4%	30.4%	49.9%
Respiratory	14.3%	9.5%	13.2%	45.7%
Digestive	10.0%	10.5%	9.6%	48.9%
Musculoskeletal	4.8%	11.0%	4.6%	51.3%
Injury	6.8%	12.8%	7.2%	54.6%
Genitourinary	5.2%	10.5%	5.3%	45.6%
Other	17.5%	11.6%	17.8%	50.8%
Race				
African American	19.3%	11.0%	19.2%	49.6%
Caucasian	76.4%	11.3%	76.8%	50.3%
Other	4.3%	9.6%	4.1%	45.9%
Gender				
Female	53.6%	11.2%	53.5%	50.1%
Male	46.4%	11.2%	46.5%	49.9%
Age				
<30 years	2.8%	11.0%	2.8%	49.9%
30–60 years	31.8%	9.9%	30.1%	46.5%
>60 years	65.4%	11.8%	67.1%	51.6%
Admission type				
Emergency	52.7%	11.4%	53.2%	50.8%
Urgent	18.4%	11.2%	18.4%	50.4%
Elective	18.2%	10.7%	17.9%	48.7%
Other	10.6%	10.7%	28.9%	49.4%
Discharge disposition				
to home	73.8%	9.4%	68.7%	45.1%
to short-term hospital	2.6%	16.2%	3.1%	60.1%
to SNF [2]	17.1%	14.6%	19.4%	57.5%
to IRF [3]	2.5%	27.9%	4.2%	73.7%

[1] Columns 2 and 3 are prior to preprocessing, column 4 is subsequent, and column 5 investigates patients with readmission over the totals after preprocessing. [2] Special Nursing Facility (SNF). [3] Inpatient Rehabilitation Facility (IRF).

3.2. Transformer-Based Prediction Model

The Transformer is a neural network architecture first introduced in 2017 by A. Vaswani et al. [21]. Their primary objective was aimed at NLP (Natural Language Processing), but the potential of this algorithm soon widened to numerous applications, as demonstrated in the present work.

The advanced architecture of the Transformer is capable of processing text or other inputs by breaking them down and analyzing the relationships between the elements involved. Its structure is presented and detailed in [21].

The mechanism has an encoder–decoder structure [22] (Figure 1). The encoder transforms the input sequence of symbols into compact and continuous representations known as context vectors or contextual embeddings. With these vectors, the decoder generates the output sequence for each element, also relying on previously generated outputs. As shown in Figure 1, the encoder is composed of six identical layers, each with two sub-layers, featuring residual connections and normalization layers. The first sub-layer contains the multi-head attention mechanism, while the second is a position-wise feed-forward network. The decoder introduces a third sub-layer, which applies an attention mechanism focused on

the encoder's output. This masking sub-layer ensures that the model does not predict the result of an element based on the influence of results from later positions in the sequence, restricting attention to elements in preceding positions only.

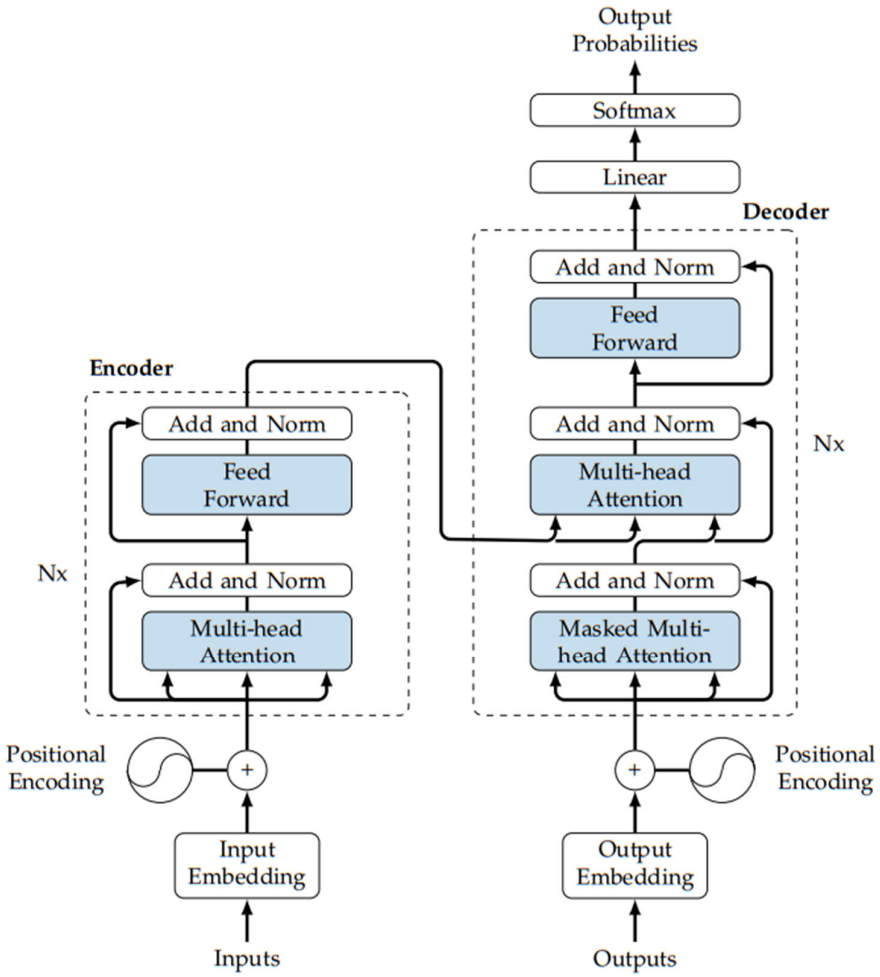

Figure 1. Architecture of the Transformer-based model.

The attention function maps a preconfigured input of queries, which are the elements that enable the attention mechanism to identify relevant information. This set of query vectors (Q) is packed into a matrix that defines the attention function. The elements are assigned a key (K), which acts as a label, and a value (V), with the actual information. Scaled dot-product attention is a method that computes the output using the dot product, which can be highly efficient for large key dimensions (d_k) compared to other algorithms such as additive attention [23]. The output matrix is as follows:

$$\text{Attention}\ (Q,\ K,\ V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The attention mechanism of a particular head follows the scheme represented in Equation (1). The query and key vectors are multiplied using the dot product, after a preliminary linearization of the Q, K, and V spaces with their respective weight matrices. These matrices allow the transformation of the input sequence into sets of vectors and are self-adjusted as the model learns during training. The initial dot product generates a set of scores that define the attention relationship between the elements of the sequence.

The obtained scores are normalized and scaled considering the dimension of the key vector to constrain and stabilize the gradients. A gradient with extreme values can

negatively affect training. SoftMax is an algorithm that generates probabilities from the scaled scores. These probabilities can be associated with the actual value of the element, which allows for assigning attention scores to these values. The obtained scores are finally combined with the values through another dot product, resulting in weighted values, which are summed to produce the output for each query vector.

Transformers simultaneously integrate as many attention mechanisms as there are execution heads available. Finally, the model concatenates the outputs of all attention heads to obtain the definitive linearized output. This new multi-head attention process significantly reduces execution times and allows for the calculation of all interactions between the elements in the sequence. The number of heads is fine-tuned in the experiments described in Section 4.1.

Transformer-based models provide an important step in text processing called tokenization, which the other models do not include. This mechanism is responsible for the division of a sequence into minimal units of meaning or tokens. A token could be a word, a sub-word, or even a character. This method offers a significant reduction in the size of the vocabulary into manageable pieces and therefore the easier handling of unknown words. The functioning of the tokenizer follows a first stage of training and building the vocabulary, where the most frequent tokens are identified and the relationships between them are established. Once the token's dictionary is built, the tokenizer is ready to divide and transform the input text. It recognizes the complete words that are included in the vocabulary and considers them a unique token, whereas it divides unknown words into sub-word tokens, trying to find hidden relationships. For example, the model may not recognize a concrete date in the 'dd-mm-yyyy' format of a medical record, so it proceeds to split the date into the biggest tokens found in the vocabulary, obtaining, for example, 'dd', '-mm', and '-yyyy'. This provides the model with temporal information of the exact day, month, and year, coming from a complete date. This example can be generalized to medical codes or other clinical information.

Embedding is a technique that allows the capturing and expressing of the main information of a sequence element in the form of a vector. In this way, the model constructs a vector space where vectors representing similar words are positioned closer together, thereby mapping a vector space of semantic and syntactic fields, which is later used to interpret new words or elements in the sequence. This component of the structure is regulated by the embedding dimension, a hyperparameter that provides the algorithm with a vectorial semantic field of a specific dimension. The dimension of this vector is fine-tuned in Section 4.1.

The intermediate dimension refers to the dimension of the intermediate state vectors for each layer. This can also be understood as the size of the hidden layers generated by the feed-forward component, and it is essential to adjust the model to the difficulty in recognizing patterns through this operation. This hyperparameter is fine-tuned in the experiments of Section 4.1.

Positional encoding is used to preserve the order of the elements in the sequence due to the absence of recurrence or convolutions in the model's structure. This solution addresses the complications of other positioning methods, such as absolute or normalized positioning, by utilizing a sinusoidal ordering where the frequency varies according to the component of the positional vector in binary encoding.

The code structure is divided into different sections. The first block separates the sequences into feature-class pairs and splits the entire dataset into training pairs (70%: 14,241 and 18,693 entries for two and three classes, respectively), validation pairs, and final test pairs (15% each: 3561 and 4005 entries for two and three classes, respectively). The second section contains the functions responsible for preparing and building the

dataset using a tokenizer and batch preprocessing. It also includes subsequent feature selection. In the final block, the Transformer model is trained using a 5-fold cross-validation strategy: training the model with 4 folds and evaluating it with the remaining fold. The Transformer is initially trained for 50 epochs using a 0.0001 learning rate and a batch size of 64. These hyperparameters are fine-tuned in the experiments of Section 4.1. The Adaptive Moment Estimation (Adam) optimizer is used to update the model weights. The Adam algorithm combines the advantages of the RMSprop and Momentum algorithms to improve the learning process of a model, reducing the convergence time. When considering a classification task with multiple classes, the Cross Entropy Loss with a Softmax layer is used to estimate the model error. The training tasks were performed with a GPU NVIDIA GeForce RTX 3060, (NVIDIA Corporation, Santa Clara, CA, USA), installed on a personal computer.

3.3. Feature Selection Methods

An important part of this work has been focused on selecting the features that optimize or worsen the model's predictive performance, using two techniques.

The first technique is the SelectFromModel (SFM) algorithm, a function imported from the sklearn.feature_selection package that selects features that meet a relevance requirement in the model's estimation. For this, a simple machine learning model is trained beforehand to obtain a weight matrix with the associated importance of each feature and then filter using a threshold or a fixed number of features. In this case, Random Forest (RF) was employed in the selection process. SFM transforms a dataset by selecting features based on their importance weights. Since the data consist of text sequences when working with a Transformer model, the data types are extracted and encoded using OneHotEncoder and LabelEncoder. This is required when using Random Forest or SVM algorithms but not when using Transformer. After determining the weights with RF, the code collects the weights of the values belonging to the same feature and takes the maximum weight, which is then normalized. This process results in a reasonably sized matrix of columns with their respective weights, which is used as the input to the SFM. The result of threshold-based selection using SFM is then decoded back into a text sequence to adapt it to the Transformer again.

The second technique is permutation feature elimination: here, we trained the model by removing features one by one (or grouped by relation), observing the accuracy evolution, and repeating this process for each individual feature (or in groups).

3.4. Methodology of Evaluation and Metrics

The objective of the experiments was to evaluate the accuracy of the model and to correct or optimize its structure to improve its prediction accuracy. A cross-validation strategy was used to develop and evaluate the system: 85% of the data were randomly selected to train and validate the model during the experiments, while the remaining 15% were reserved to assess the model's accuracy with data not seen during system development. In this way, a real-world situation was modeled in which the patient may have had a combination of characteristics that the model had not encountered before. The training data were divided into folds to execute a 5-fold cross-validation. In each training iteration, four of the folds served to train the Transformer model and the remaining fold was used to validate the training, with us ensuring that adjustments and evaluations were made with different combinations of data, as shown in Figure 2. This process was repeated 5 times and the evaluation metrics were the averages taken from the 5 iterations.

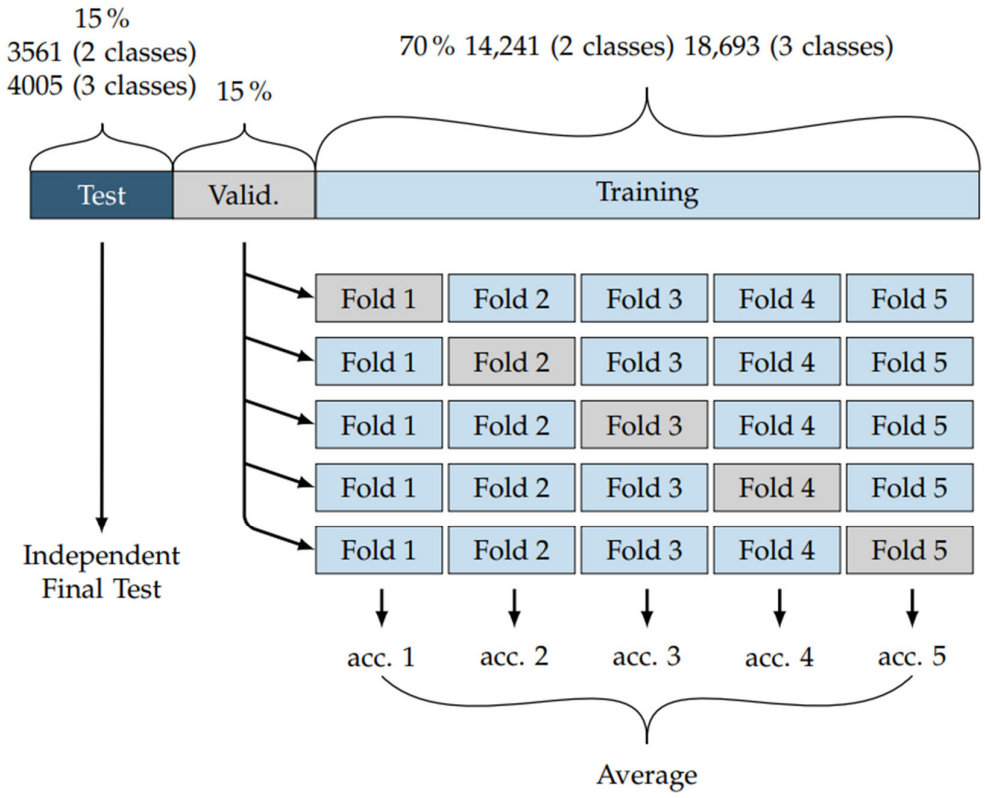

Figure 2. Distribution of the dataset into training (70%), validation (15%), and test (15%) pairs. Visualization of the cross-validation process with 5 folds.

The model's performance was evaluated using the accuracy (*acc*) and confidence interval (*CI*) with a 95% level of significance (Equation (2)). Below, N is the number of test pairs.

$$CI = \pm 1.96 \cdot \sqrt{\frac{acc \times (100 - acc)}{N - 1}} \quad (2)$$

Classification performance was also examined using confusion matrices. The rows represent the actual classes, while the columns represent the classes estimated by the model. The main diagonal, consisting of true negatives (TN) and true positives (TP), reflects the accuracy of the prediction, while the other cells explain the type of confusion. A perfect model would only have samples in the main diagonal. The classes are expressed as 0, 1, and 2 for NO, LESS, and MORE, respectively.

One way to measure these relationships is through the false positive rate (FPR), sensitivity or true positive rate (TPR), and specificity or true negative rate. These two statistical variables, TPR and FPR, can be embodied in a receiver operating characteristic (ROC) curve to obtain a visual representation of the model's performance. The closer the curve is to a null value of FPR, the better the model will detect true negative cases. The area under the ROC curve (AUC) is an estimator of the model's ability to distinguish between classes, also serving as a summary of the ROC curve. A random classification results in an AUC value of approximately 0.5.

4. Experiments and Discussion

The experiments were conducted following two main analyses: hyperparameter tuning and feature selection. The primary evaluation metrics were accuracy and AUC, which allowed for a comparison with previous studies. The planning of each experiment when designing the next one was based on the results of the accumulated accuracy and the

AUC, and never on the final test accuracy. The final test accuracy is reported at the end of this Section with the best system configuration (obtained during the system development).

4.1. Hyperparameter Tuning

Hyperparameters are adjustable variables that control the learning process of the Transformer-based model. Tuning a single hyperparameter can be a straightforward and quick task, but it becomes considerably slower as more hyperparameters are considered. The process consisted of an initial coarse adjustment of hyperparameters, followed by a second manual tuning that adhered to a predetermined order based on which parameters have the most impact on the learning rate: learning rate (LR), epochs, batch size, embedding dimension, intermediate dimension, and number of heads.

The accuracy showed a favorable evolution (Table 2). The confidence intervals did not overlap between the initial trials (around 58.64% for two classes) and those with fine-tuned hyperparameters. When compared with the results from the reference article [18], we found that the Transformer-based model presented in this work achieved a better accuracy than all other proposed algorithms.

Table 2. Best accuracy results after hyperparameter tuning for two and three classes.

Experiment	LR	Epochs	Batch Size	Embed. Dim.	Interm. Dim.	Num. Heads	AUC	Accuracy (%)	F-Score (%)
Fine-tuned 2 classes	6.25×10^{-5}	30	128	192	2048	20	0.619	61.98 ± 0.75	61.88 ± 0.75
Fine-tuned 3 classes	6.25×10^{-5}	30	128	192	2048	16	-	44.78 ± 0.77	44.40 ± 0.77

The learning rate (Figure 3a) is often the hyperparameter that most influences the model's performance. It determines the ability of new information acquired by the model to override previous knowledge and controls the response during learning in relation to the estimated error each time the model weights are updated. In this study, a more advanced configuration was trialed using a scheduler. The scheduler maintained an initial learning rate (LR = 6.25×10^{-5}), obtained during the tuning process, which then decreased after the first five epochs by a factor of 0.905 in each iteration. However, while the confidence intervals overlapped, making it difficult to determine with certainty, it appeared that the scheduler did not significantly improve the accuracy or the AUC during training and validation. With that said, the selected hyperparameters offered the best results for the validation subset. Beyond the learning rate, another considered hyperparameter was the number of epochs (Figure 3b), which is the total number of complete iterations of a process. This number must be set appropriately (30 in this study), with a minimum to ensure the model's convergence and an upper limit to avoid overfitting. Furthermore, regarding the batch size, the best performance was obtained for 128, meaning we used this value for further experiments.

The embedding dimension of the Transformer (Figure 3d) showed an oscillating behavior. Although the differences were not significant, the value with the best result (192) was considered. Moreover, regarding the number of heads (Figure 3e), good performance was obtained for values between 16 and 20 heads. Increasing the number of heads augmented the model complexity, requiring more data and meaning the accuracy decreased. Accordingly, in this work, the value of 16 was selected to have a simpler model with a smaller number of parameters to train. Finally, for the intermedia dimension between the encoder and the decoder, a value of 2048 (Figure 3f) was considered as a good compromise between model complexity and the amount of data. The best accuracy is obtained when the model complexity is sufficient for learning from the available amount of data: on the

one hand, a higher-complexity model cannot be trained properly, producing worse results; on the other hand, a low-complexity model cannot execute the specific task.

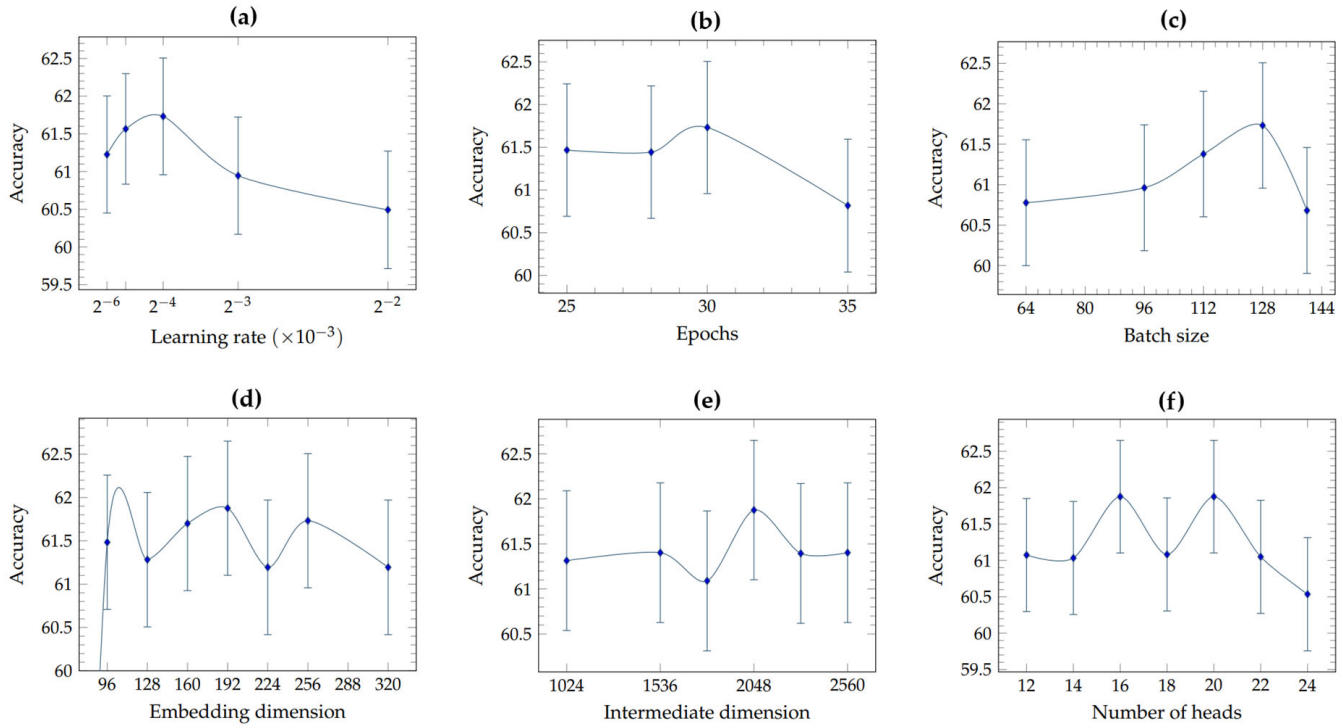

Figure 3. Evolution of the accuracy of the model with (**a**) learning rate; (**b**) epochs; (**c**) batch size; (**d**) embedding dimension; (**e**) number of heads; (**f**) intermediate dimension.

4.2. Feature Selection

The second block of experiments was focused on feature selection, where the most relevant information in this classification task was identified. The aim of our elimination was to enhance medical procedures, reducing the complexity (number of features considered) and maintaining the same accuracy. The model was optimized by reducing the number of features through two methods:

4.2.1. SelectFromModel (SFM)

For the selection of features using SFM, a threshold was considered based on the quantile of the weight matrix. This matrix was built with the maximum weight associated with the values for each feature. The quantile was introduced as a parameter in the SFM function to transform and reduce the dataset. The best results were obtained when using the third quantile threshold.

Table 3 summarizes the results after feature selection using the SFM method with the third quantile threshold.

Table 3. Accuracy after feature selection using the SFM method with the third quantile threshold.

Experiment	AUC	Accuracy (%)	F-Score (%)
SFM for 2 classes	0.620	61.96 ± 0.75	61.82 ± 0.75
SFM for 3 classes	-	44.18 ± 0.77	42.83 ± 0.77

The weights of each feature were gathered in a separate file during 5 runs, and then they were averaged. A matrix was obtained using the feature importance attribute of the Random Forest classifier, which calculates the importance of each variable as the mean and standard deviation of the accumulation of impurity reduction within each tree in

the forest [24]. However, these weights can be confusing and deceptive in the case of high-cardinality variables, creating the need for statistical techniques capable of estimating the actual contribution of each feature to the trained model. This study considered two techniques. The first one was the permutation of the variables, one by one, aiming to break the relationship between a particular feature and the target class, as detailed in the following Section. The second method involved studying the variability among the values within the same feature using the coefficient of variation (CV), a simple indicator composed of the mean μ and standard deviation σ, which provides insight into the homogeneity of a dataset and the extent of fluctuations among the values of a patient variable.

$$CV = \frac{\sigma}{\mu} \tag{3}$$

Variables with higher importance weights, but also presenting a significant variation between their different values, had a greater impact on the accuracy (see Figure 4). This affected features such as gender (2), which, despite having a high maximum importance score, did not contribute much to determining readmission (its two values showed almost no difference). Two variables stood out above the rest in this regard: discharge disposition (5) and number of inpatient visits (13).

The variables with the least influence in the model's classification were those related to pharmacological treatments. Within this group, the most significant were insulin (35), change in medication (41), glipizide (26), and the number of drugs administered (10). Insulin is closely linked to type 1 diabetes (DM1), indicating a smaller group of patients with more severe episodes. Insulin is also associated with therapeutic failure in type 2 diabetes (DM2). Of the total database, 58.12% of readmitted patients had been administered insulin in varying doses, and of those with high doses, 55.15% were readmitted (Table 1). It is important to note that once the data are balanced, variations close to 50% may reflect much larger real-world discrepancies in the clinical series, and they can be quite significant. As the least significant variables in this group, treatments involving tolbutamide (28), troglitazone (33), tolazamide (34), glipizide–metformin (37), glimepiride–pioglitazone (38), and metformin–rosiglitazone showed near-zero importance, likely due to a lack of positive samples for these drugs.

The primary and secondary diagnoses (14, 15, 16) held a variability of more than 50% in CV. Genitourinary diseases, especially nephropathies and those related to the filtration of high blood glucose levels; circulatory and vascular conditions; digestive diseases; and tissue injuries, particularly necrosis, were among the most relevant factors for estimating a patient's likelihood of readmission. In contrast, tumors and neoplasms showed a weaker relationship with diabetes and readmission due to this condition.

The analyses showed that the HbA1c test result (19) may be more relevant than the blood glucose serology test (18) in tracking the patient and predicting a possible relapse in the severity of the disease. This aligns with several medical guidelines [25,26], which opt for the HbA1c test as a diagnostic and monitoring measure. The number of diagnoses (17) ranked as the third most relevant variable, with the strongest relationship between the maximum weight and CV, acting as a significant indicator of the patient's disease complexity, as well as heavily influencing the DRG value, which explains its high importance weight.

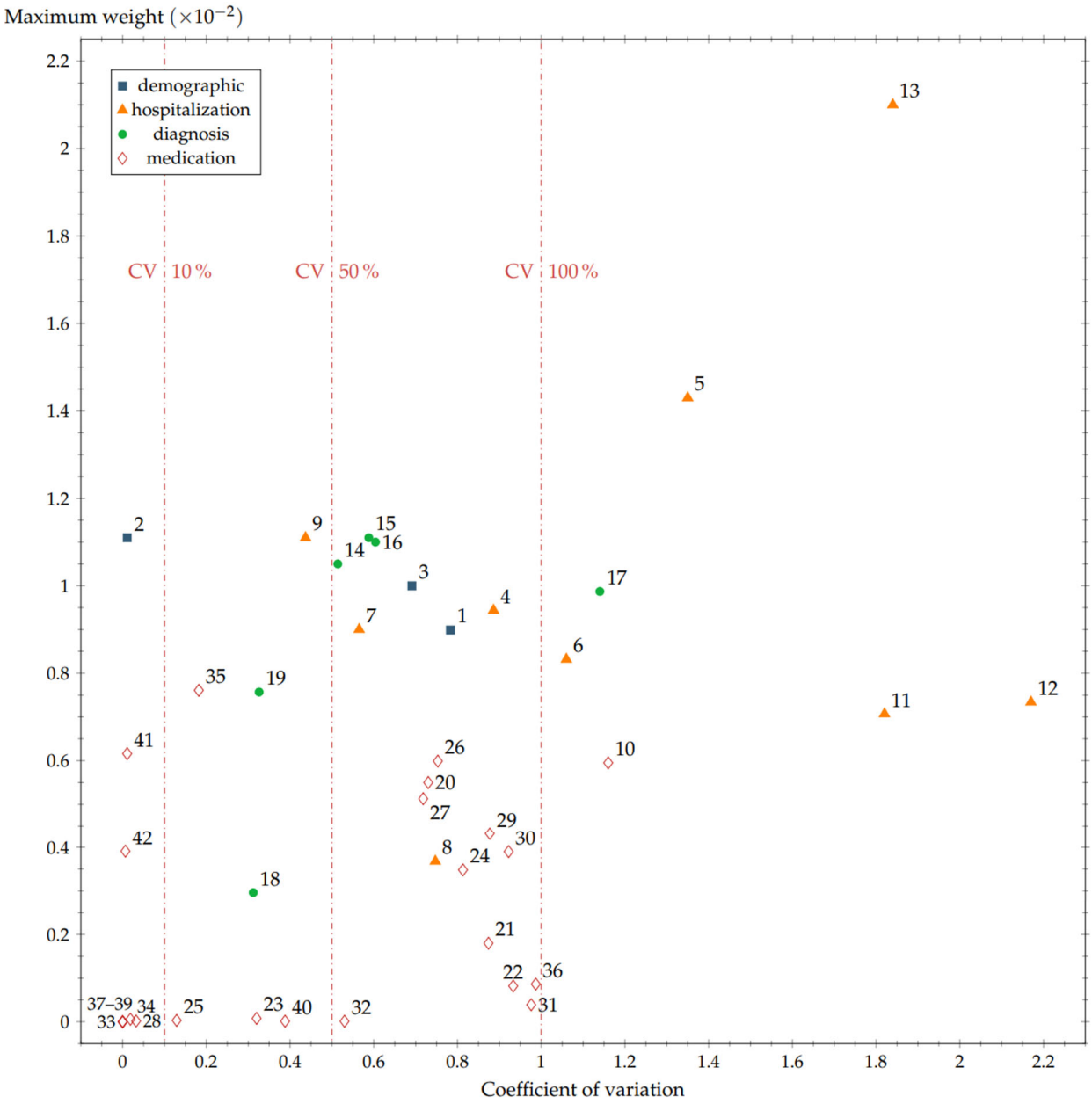

Figure 4. Maximum of the weights of the values of each feature compared to the corresponding CV. Feature list: 1 race; 2 gender; 3 age; 4 admission type; 5 discharge disposition; 6 admission source; 7 time in hospital; 8 number of lab procedures; 9 number of procedures; 10 number of medications; 11 number of outpatients visits; 12 number of emergency visits; 13 number of inpatient visits; 14 diagnosis 1; 15 diagnosis 2; 16 diagnosis 3; 17 number of diagnoses; 18 glucose serum test result; 19 A1c test result; 20 metformin; 21 repaglinide; 22 nateglinide; 23 chlorpropamide; 24 glimepiride; 25 acetohexamide; 26 glipizide; 27 glyburide; 28 tolbutamide; 29 pioglitazone; 30 rosiglitazone; 31 acarbose; 32 miglitol; 33 troglitazone; 34 tolazamide; 35 insulin; 36 glyburide–metformin; 37 glipizide–metformin; 38 glimepiride–pioglitazone; 39 metformin–rosiglitazone; 40 metformin–pioglitazone; 41 change in medications; 42 diabetes medications.

When examining the group of demographic variables, we observed a strong predictive performance of the age group (3). Although gender (2) had the highest weight among the three characteristics, its low CV suggested that it does not significantly contribute to the prediction of readmission. Age, on the other hand, displayed a reasonable increase in weight across age intervals from younger to older (with a milder effect observed in the 80–90 years group, and with no effect in the 90–100 years group, which exceeded the life

expectancy during the dataset's timeframe). Additionally, the model identified Caucasian and African American races (1) as the most likely to be readmitted, although some bias may have arisen due to data disparities between ethnicities.

The features providing information regarding a patient's hospitalization generally had the most significant value in predicting the readmission of a patient with diabetes mellitus. Among these, the most important was the number of inpatient visits of the patient in the year preceding the encounter (13). The most determinant value in this column corresponded to zero previous hospitalizations, which accounted for 60.39% of the processed database samples, followed by patients admitted once (20.59% of the samples). The remaining approximately 20% consisted of patients who had been readmitted more than once previously.

Figure 5 illustrates the relationship between annual hospitalization history and the percentage of readmissions within that group. The red curve shows a clear proportionality of the probability of readmission with the evidence of prior admissions, except for the last values, which may have differed due to the scarcity of examples. This relationship contrasts well with the green line, which represents the normalized weights from the average of five data samples. While the higher values exhibit considerable uncertainty and are not very reliable, the lower values (0–3 previous hospitalizations) demonstrate that the model makes effective predictions based on whether the patient has been admitted before.

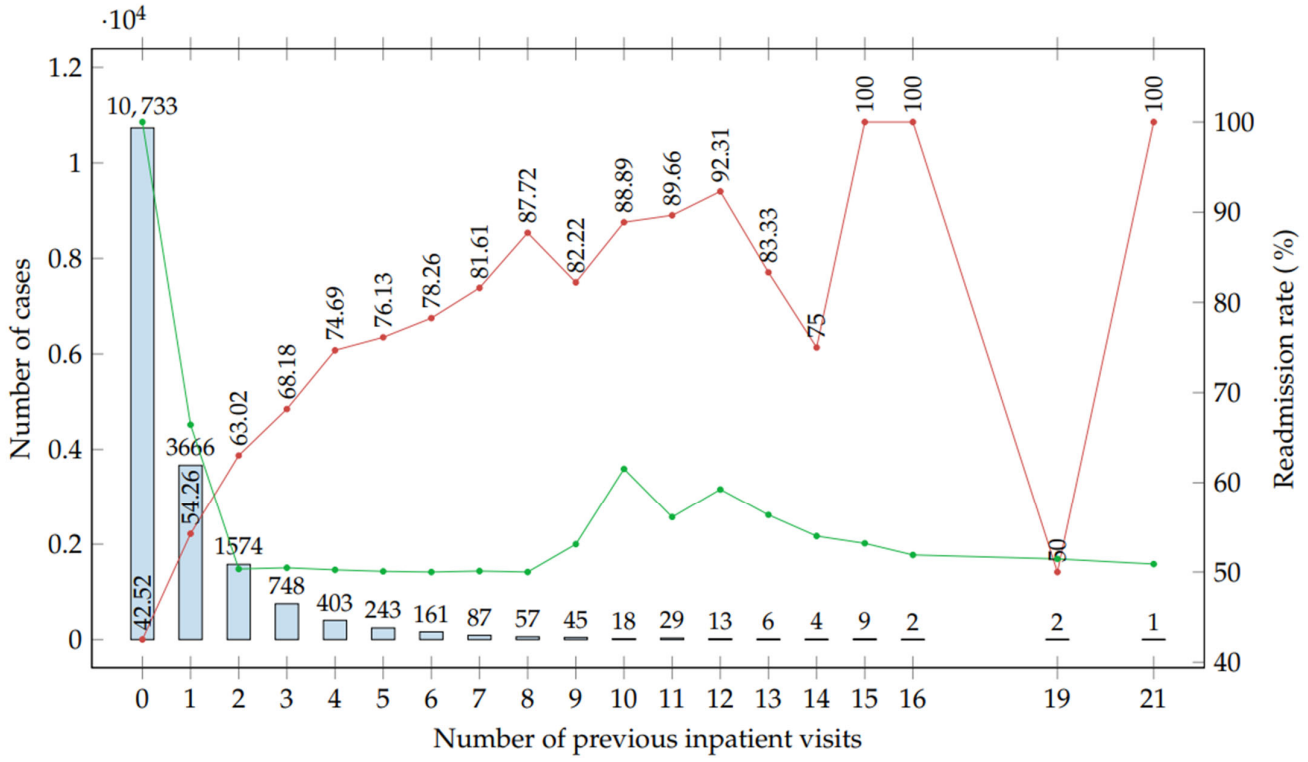

Figure 5. Patient inpatient visits in the preceding year (13) and readmission rate for each feature value (red line). Relevance weights are presented in green.

The discharge disposition (5) was the second feature with the most impact, and we found it had a CV of over 100%. In the literature, it is recognized as an independent predictor of readmission for other diseases [27], with routine discharge standing out as the most relevant type of discharge. Other factors linked to hospitalization that considerably influenced the model's learning included the number of lab tests performed during the stay (9) and the reason for admission (4), where emergency and urgent admissions were particularly notable. Additionally, the length of stay (7) played a role, with the model

giving more weight to patients who stayed only one day or beyond the eighth day when classifying potential readmissions.

4.2.2. Permutation Feature Elimination

A greater variability in the accuracy when eliminating a feature indicated it had a greater impact, whether favorable or adverse, on the learning process. A feature-by-feature elimination enabled a permutation based on the properties of diabetic patients to corroborate the results obtained using the previous technique, SFM. Figure 6 illustrates the performance of the Transformer-based model, measured according to its accuracy without each feature.

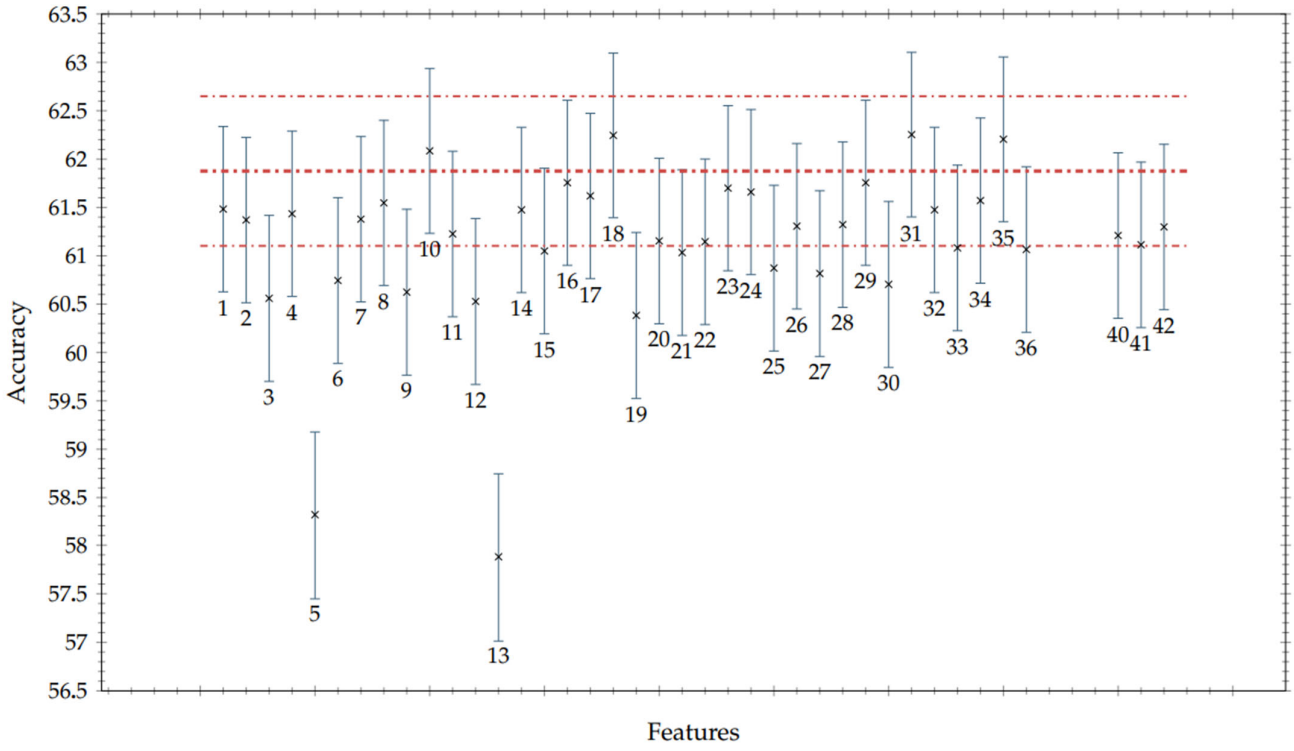

Figure 6. Accuracy results in removing features by permutation. Red dashed lines mark the best accuracy obtained in the prior stage of hyperparameter tuning, with its CI bounds.

The results in Figure 6 demonstrate the coherence between this feature selection technique and the previous one. When the discharge disposition (5) and the number of inpatient visits (13) were removed, the model's accuracy dropped significantly below 58.5%, confirming that the algorithm pays more attention to these variables. Many of the columns fell within the confidence interval of the best hyperparameter tuning result, indicated in red. The model achieved maximum accuracies of 62.25% ($\pm$0.75) and 62.20% ($\pm$0.75) without acarbose (31) or insulin (35), respectively. Insulin is a clear example of how a feature to which the model assigns a relatively high importance weight (more than any other medication) does not actually contribute to the final class prediction because it shows very low variability (CV < 10%). In addition to sequentially eliminating individual features, complete groups were also removed. Dropping the group of variables related to the hospitalization stay significantly reduced the accuracy to 54.87% (Table 4).

Table 4. Accuracy after feature and group elimination by permutation (a dashed line separates the isolated feature results from group-of-features results).

Feature/Group Removed	AUC	Accuracy (%)	F-Score (%)
Acarbose	0.623	62.25 ± 0.75	62.15 ± 0.75
Blood glucose serologic test	0.622	62.24 ± 0.75	62.05 ± 0.75
Insulin	0.622	62.24 ± 0.75	62.04 ± 0.75
Demographic group	0.604	60.40 ± 0.75	60.39 ± 0.75
Hospitalization group	0.549	54.87 ± 0.76	54.86 ± 0.76
Diagnostic group	0.617	61.68 ± 0.75	61.34 ± 0.75
Medication group	0.613	61.32 ± 0.75	61.28 ± 0.75

4.3. Experiments with Selected Features

After achieving initial success with the fine-tuning experiments, certain limitations were found during the second stage of feature selection. H. Lu and S. Uddin [18] accomplished better results in some models during feature selection, even when starting from a lower initial accuracy. In their SVM model, an absolute improvement of over 10% was reported when removing the features. This result surprised us because the rest of the methods, including our Transformer-based model, did not produce such an improvement (when using SFM) in a binary classification. To analyze this further, we then decided to evaluate the SVM algorithm and train it under the same working environment as the Transformer-based model. This experiment allowed for a direct comparison within the same practical framework, using an identical code, methods, and dataset.

SVM is a machine learning algorithm with an architecture designed to handle a quadratic optimization problem by finding a hyperplane that separates the classes and maximizes the margin between the closest point of each class and the hyperplane itself. To achieve this, SVMs rely on support vectors, a subset of the data that, during training, define the margins and thus the position of the classification hyperplane.

The designed code follows the same structure as the Transformer-based model, importing the model from the Scikit-Learn library and excluding unnecessary functions such as the tokenizer and batch preprocessing. The implementation also incorporates a combination of encoders—OneHotEncoder, OrdinalEncoder, and LabelEncoder—to adapt the input data for the SVM.

However, when applying the SFM mechanism to the SVM model developed in this work, the accuracy only improved by 0.11% (Table 5). This result can have several explanations. First, the model designed by H. Lu and S. Uddin [18] may implement a different feature selection strategy. Although the authors mention the use of SFM, it has multiple applications and parameters that influence how features are chosen. Additionally, the structure and design of the SVM model code differ between works. There are many ways to fine-tune the classifier, as well as a variety of libraries available to import the algorithm, and several encoders to apply to the dataset. However, the accuracy without feature selection was more comparable, suggesting that the issue likely lay in the selection process itself. Furthermore, differences in preprocessing, mainly due to a lack of detailed information or the codes used for the features, may also have affected accuracy. Finally, the Random Forest model was also applied, yielding accuracy results of 61.13% and 61.26% before and after feature selection. Additionally, a comparison of the computational resources required between models was performed. While Transformer-based models achieved a better predictive performance than traditional machine learning models, such as SVM or RF, their neural network nature demanded sophisticated computational resources, such as GPUs, to attain reasonable processing times. In this model, each epoch during training took 21 s, with step durations ranging from 164 to 178 milliseconds. Testing with 15% of

the samples lasted between 150 and 240 s. The model required approximately 15 min to compute one cross-validation fold, including the training, validation, and test processes.

Table 5. Accuracy of SVM model.

Experiment	AUC	Accuracy (%)	F-Score (%)
SVM without feature selection	0.613	61.24 ± 0.75	61.00 ± 0.75
SVM with feature selection by SFM	0.614	61.35 ± 0.75	61.12 ± 0.75

In order to achieve better results, we delved deeper into the classification mechanism of this particular model. The Transformer-based algorithm aims to focus attention on relevant features to generate a response. In the case of binary classification, this solution involves classifying whether a patient will be readmitted within 30 days (LESS) or not. However, within the "not" category, there are two sub-classes, NO and MORE, and incompatibilities between them potentially confuse the model. When analyzing the confusion matrix for a multiclass classification problem, such as that in Figure 7, the false rate elements with the highest values were the predictions of MORE classed as LESS (12.12%). The NO–MORE confusions, in both directions, accounted for more than 35% of the total misclassifications. This pattern was consistent across most multiclass confusion matrices, indicating that it might be a good idea to classify a readmission that occurs later than 30 days as NO rather than LESS in a binary classification. Nevertheless, the boundary between these classes was not well defined, so it is crucial to consider other combinations of samples.

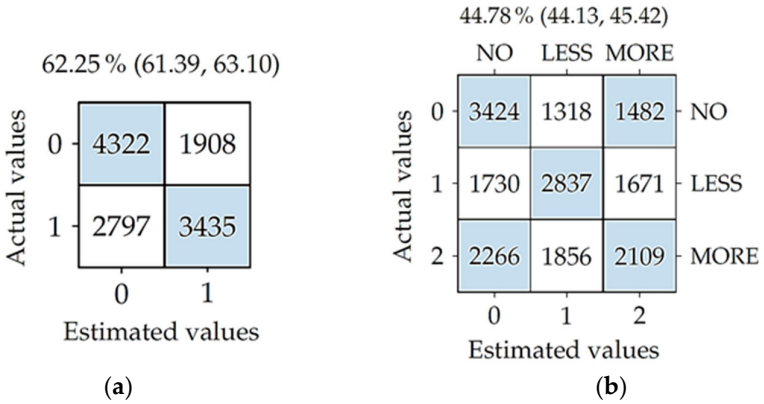

Figure 7. Confusion matrices of the best accuracy results for (**a**) binary and (**b**) multiclass problems.

The model was significantly better at discriminating between the classes NO and LESS than when an intermediate phase was included, such as the MORE class. With this configuration, the cumulative accuracy reached 65.82% (Table 6), with a confidence interval that overlapped by 0.04% with the best result when combining the NO and MORE classes. Nevertheless, the algorithm struggled with distinguishing between the LESS and MORE classes. Adapting the model's learning process to these results could be a promising option. The model tended to predict more false negatives than false positives overall. Although this relationship decreased with three classes, as more weight was distributed toward the MORE class, the model remained more specific than sensitive. In a medical context, this is a flaw, and work on the algorithm's design should focus on reversing this situation, as it is preferable to "prevent rather than cure".

Table 6. Prediction experiments with different combinations of classes.

Experiment	AUC	Accuracy (%)	F1-Score (%)
LESS vs. NO	0.655	65.55 ± 0.74	65.51 ± 0.74
LESS vs. NO + feature selection by SFM	0.657	65.82 ± 0.74	65.71 ± 0.74
LESS vs. MORE	0.589	58.94 ± 0.76	58.82 ± 0.76
MORE vs. NO (56,480 encounters)	0.612	61.25 ± 0.77	61.23 ± 0.77

4.4. Comparison of Machine Learning Algorithms

This Section compares different machine learning algorithms.

Tables 7 and 8 summarize the results obtained with different machine learning algorithms, for investigations including all features and those when selecting the features using SFM, respectively. As shown, the proposed method (based on Transformers) can slightly improve the performance obtained with conventional machine learning algorithms. Although the difference is not statistically significant, our proposal can obtain good results without preprocessing modules required to handle different types of information. When using all features (Table 7), the improvements versus SVM and RF reported in [18] are statistically significant.

Table 7. Prediction experiments considering all features.

Experiment: All Features Model	AUC	Accuracy (%)	F1-Score (%)
SVM	0.613	61.24 ± 0.75	61.00 ± 0.75
SVM [18]	-	57.61 ± 0.76	-
RandomForest	0.612	61.13 ± 0.75	60.95 ± 0.75
RandomForest [18]	-	60.35 ± 0.76	-
Transformer-based system (proposed)	0.619	61.98 ± 0.75	61.88 ± 0.75

Table 8. Prediction experiments using the same number of features selected using SFM.

Experiment: Feature Selection by SFM Model	AUC	Accuracy (%)	F1-Score (%)
SVM	0.614	61.35 ± 0.75	61.12 ± 0.75
RandomForest	0.612	61.26 ± 0.75	61.07 ± 0.75
Transformer-based system (proposed)	0.620	61.96 ± 0.75	61.82 ± 0.75

4.5. Global Results

Test pairs were set aside from the training dataset to evaluate the model's performance with a new group of diabetic patients when predicting their readmission. To this end, 15% of the samples were reserved, with the final class estimated after the algorithm had completed its training. It is important to consider that the confidence intervals of the final test were slightly higher than the mean of those from the training data, as the test data segment contained fewer samples. The accuracy in the final test stood out when eliminating race (62.96 ± 0.75%). With three classes, fewer experiments were conducted. This was partly because it was unnecessary to duplicate tests for hyperparameter tuning and other processes since they are developed under the same architecture. The model achieved up to 44.78% accuracy during training and validation, and 46.47% in the final test. For both type of problems, binary and multiclass, the confidence intervals of the initial tests and the best results did not overlap, indicating a reliable improvement in the accuracy. Another relevant finding, considering that the samples in the database were balanced, is that the model was overall best at classifying the records of patients who were not readmitted, followed

by those readmitted in less than 30 days, and, finally, the MORE class (readmission after 30 days).

4.6. A Case Study: Defining a Protocol for Diabetic Patients

Inpatient care units place a significant emphasis on and put great effort into programs that mitigate the rate of potentially avoidable readmissions. Moreover, the rate of unpredicted readmissions within 30 days after discharge is considered a quality metric of the effectiveness of healthcare for many hospital systems [28,29]. To avoid readmissions, clinicians usually draw on their expertise and prediction techniques that help to estimate the severity of a certain disease. Nevertheless, inpatient care units usually receive patients suffering comorbidities and complex diseases, and more sophisticated methods of prediction are thus required for readmission reduction. In this situation, artificial intelligence algorithms can help to analyze multivariable problems, providing useful information for clinicians. This paper proposes a Transformer-based algorithm for predicting readmissions, and we performed a detailed analysis of the main features in this estimation.

This Section describes the process of improving a clinical protocol based on this work's finding. Professionals from several hospitals in Madrid, Spain, were consulted, to verify the coherence of the results of the feature analysis.

Unlike other diseases, there is no established standard protocol for managing diabetes, and it is strongly recommended that a structured discharge plan should be tailored to the individual patient with diabetes [30,31]. Accordingly, a protocol for discharging patients with diabetes mellitus is proposed, which was developed in collaboration with clinicians. This incorporates a prediction model in several steps of the inpatient-to-outpatient procedure, which ensures consistency and evidence-based care throughout the patient's treatment journey. Figure 8 shows a flow diagram of the proposed protocol.

The first stages of the transfer process often comprise the stabilization of glycemic levels and gaining cardiovascular control of the patient (Figure 8, stage 1). This is subsequent to the patient's hospitalization, which has a described actuation plan [32]. In discharge, the HbA1c results (19) are often considered as the main variable to determine the evolution of a diabetic patient [33], together with other features such as the glucose serum test results (18) and other risk factors such as the discharge status (5), comorbidity index, or educational status [34].

Diabetic patient admissions from emergency departments are estimated to comprise between 11% and 21% of all diabetic outpatients [35]. In patients with diabetes, readmission rates range between approximately 14% and 23% [34], depending on many factors, some of them described later. In the imbalanced dataset used for this study, only 53.91% of the hospitalized patients who stayed between 1 and 14 days were not readmitted, while 46.06% experienced a readmission (whether avoidable or not), with 11.16% being readmitted within 30 days. These results corroborate the premise that predicting a complex disease such as diabetes presents a multivariable problem influenced by numerous risk factors. While some readmissions are unavoidable, many are preventable, and prediction models could help professionals detect these situations by adding to their expertise, thus guiding clinicians toward better solutions that match a patient's circumstances.

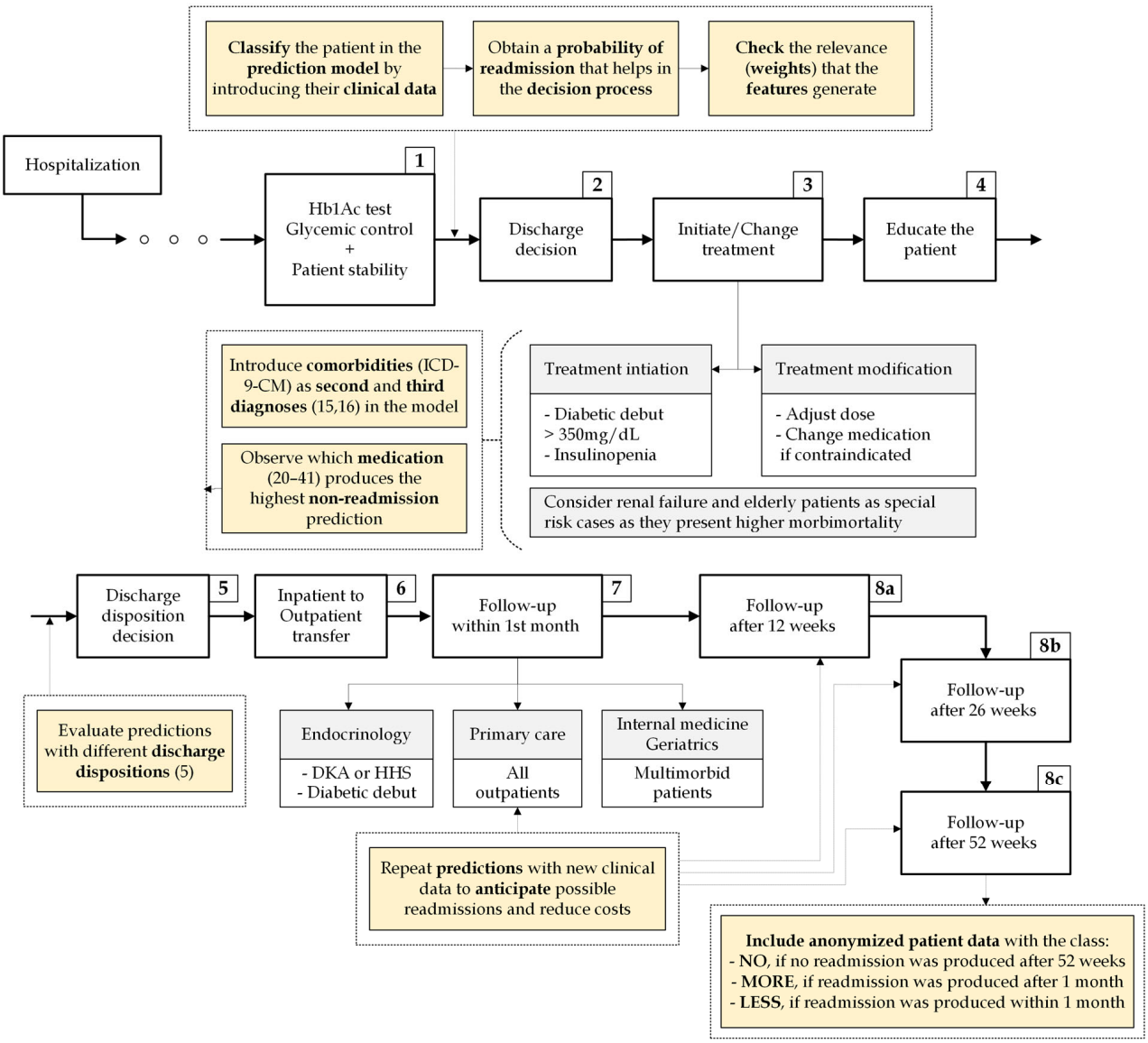

Figure 8. Protocol proposed for the discharge process of diabetic patients. Every main stage is indicated with a number. Colored boxes indicate the possible implementation of the Transformer-based model in this protocol. DKA, Diabetic KetoAcidosis. HHS, Hyperosmolar Hyperglycemic State.

The most relevant patient information found in this work—their inpatient visits (13) and discharge disposition (5)—considerably influences the readmission of diabetic hospitalized patients. Features that significantly influence the likelihood of readmission and are not established when a discharge choice must be made, such as discharge disposition (5) or time in hospital (7), can be included by the clinician in the decision process to predict a possible readmission while considering different short-term scenarios for the patient. A prediction tool used at this point should incorporate the importance of each feature in each circumstance when producing the final estimation.

Pinto et al. [35] explained that, when facing diabetic treatment (Figure 8, stage 3), three differentiated groups of patients can be distinguished: those in who initiation of a treatment is essential, those in who it is essential to think about a modification of the treatment, and those that do not have enough clinical and analytical data to give certainty about the need for a therapeutic change in their treatment at the moment of discharge. Patient readmissions are also highly dependent on disease complications and possible comorbidities developed. For example, kidney failure and elderly patients create a higher

risk of developing complications. Transformer-based models outperform in these situations, as due to their structure, they are proficient at integrating diseases and introducing new features without the need for the development of a specific model for each type of disease. Diabetes educators should provide patients and their close relatives with sufficient information about the treatment modalities and related safety considerations, particularly for insulinization (Figure 8, stage 4). This information should include instructions for injection and autoanalysis, the insulin regimen, when a modification of the dose is needed, non-insulin drugs, hypoglycemia prevention and treatment, and basic knowledge of the foods they can safely consume. Stage 5 of the protocol denotes the discharge decision, for example, home or short-term hospital (see Table 1 for the most common discharge options). The model can be used as a supplementary tool to evaluate different discharge scenarios (5), one that reinforces professional decision-making [36].

After the transfer from inpatient to outpatient, a follow-up of patients' circumstances during the first month is crucial to achieve acute care. In addition to accessing primary care follow-up, patients who underwent a diabetic debut, diabetic ketoacidosis, or a hyperosmolar hyperglycemic state during their admission must be prompted to attend an appointment with an endocrinologist. At this stage, particularly within the first month but also for further monitoring, predictions could be made based on the patient data to anticipate possible readmissions and deliver early prevention, thus reducing the aggravation of the disease and additional costs associated with a readmission.

Finally, patient data could be included with their corresponding classes to improve the model proposed, using data obtained from the inpatient care unit and follow-up services. This case study was designed in close collaboration with clinicians from different hospitals in Madrid, with the objective of implementing a real protocol where the Transformer-based model has an important role, which it fulfills with an enhanced performance in several characteristics compared to previous models. The results of the execution of this initiative could be described in future work.

5. Conclusions

This work focused on the implementation of a Transformer-based model to predict, analyze, and interpret results in estimating the readmission of patients with diabetes mellitus. Compared to other algorithms like SVM and RF employed in previous studies, Transformers achieve better predictions, measured in terms of accuracy, AUC, confusion matrices, and other indicators. This improvement was also observed after feature selection under the same conditions, reaching an accuracy of 62.25% for two classes and 44.78% for three. One of the main advantages of this model is its versatility for processing different types of inputs or features, due to the adaptability of the sequences and the use of a tokenizer. With other traditional machine learning models, such as SVM or RF, many code functions must be considered to adapt the features to the classification problem and its nature. This means that, for example, if a new dataset with novel features is introduced, the architectures of previous models need to transform and encode these variables to make the necessary adaptations. In contrast, Transformers perfectly accommodate these variations because they pay attention to the weight of each word in a sentence (sequence of features). The sentence can contain, as demonstrated in this work, information on patient encounters. This characteristic, combined with Transformers' ability to tackle multiclass problems without the need for adjustment, presents an opportunity for professionals to estimate the outcome of a patient with different characteristics or orders, perhaps because they come from another healthcare system or for other reasons. Moreover, the Transformer-based model has a great capacity for application to other diseases and to assist in medical decision-making in a broader context. The ability to unify the architectures of the models

used for different applications is significantly useful in promoting greater understanding and interpretation among professionals. However, some limitations were found in the explainability of these models compared to statistical techniques such as discriminant analysis. Feature analysis complements the absence of an explanatory capacity by providing a study of each feature contribution, coherent with Lu et al. [18] and the medical manuals of hyperglycemia management. Further explainability strategies, such as Local Interpretable Model-agnostic Explanations (LIMEs), could be integrated to decrypt predictions and build more reliable and interpretable models in clinical environments. This strategy will be considered in future work.

Although the accuracy is improved compared to other algorithms, it is necessary to continue working on and researching the optimization of accuracy and sensitivity. For future directions, several lines of research are proposed. Firstly, as Lu et al. [18] suggested, an ensemble of models using techniques like stacking or boosting, researchers should combine multiple architectures into a meta-model, including the Transformer-based architecture. The second direction is similar and involves mixing expert systems, which entails implementing a mechanism where different models or parts of models specialize in subtasks. This can be particularly useful when the model struggles to differentiate between two classes. Secondly, it is important to explore which feature values are related to each final class of readmission: within 30 days (LESS), no readmission (NO), or more than 30 days (MORE). This analysis, in addition to clarifying the reason for the patient's classification for the professional, serves an academic function, helping students identify which symptoms or variables are associated with a potential readmission. Thirdly, for situations with an imbalanced distribution of the original dataset, other balancing techniques (e.g., Synthetic Minority Over-Sampling Technique, SMOTE) should be evaluated rather than randomly under-sampling the data. Finally, the proposed approach will be evaluated with other public datasets, to gauge its generalizability.

Author Contributions: Conceptualization, M.G.-M. and R.S.-S.; methodology, J.G.-M., M.V.-M. and M.G.-M.; software, J.G.-M., M.V.-M. and M.G.-M.; validation, J.G.-M. and R.S.-S.; formal analysis, J.G.-M., M.V.-M., M.G.-M. and R.S.-S.; investigation, J.G.-M. and R.S.-S.; resources, R.S.-S.; writing—original draft preparation, J.G.-M. and R.S.-S.; writing—review and editing, J.G.-M., M.V.-M., M.G.-M. and R.S.-S.; visualization, J.G.-M. and R.S.-S.; supervision, R.S.-S.; project administration, R.S.-S.; funding acquisition, R.S.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Project ASTOUND (101071191—HORIZONEIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22), BeWord (PID2021-126061OB-C43) and TRUSTBOOST (PID2023-150584OB-C21), funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

Data Availability Statement: This study used previously recorded data that are publicly available.

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

References

1. IDF. *Diabetes Atlas*, 10th ed.; International Diabetes Federation: Brussels, Belgium, 2021; pp. 2–10.
2. Rubens, M.; Ramamoorthy, V.; Saxena, A.; McGranaghan, P.; McCormack-Granja, E. Recent trends in diabetes-associated hospitalizations in the United States. *J. Clin. Med.* **2022**, *11*, 6636. [CrossRef] [PubMed]
3. Van der Does, A.M.B.; Kneepkens, E.L.; Uitvlugt, E.B.; Jansen, S.L.; Schilder, L.; Tokmaji, G.; Wijers, S.C.; Radersma, M.; Heijnen, J.N.M.; Teunissen, P.F.A.; et al. Preventability of unplanned readmissions within 30 days of discharge. A cross-sectional, single-center study. *PLoS ONE* **2020**, *15*, e0229940. [CrossRef]
4. Van Walraven, C.; Bennet, C.; Jennings, A.; Austin, P.C.; Forster, A.J. Proportion of hospital readmissions deemed avoidable: A systematic review. *Can. Med. Assoc. J.* **2011**, *183*, e391–e402. [CrossRef] [PubMed]

5. Yadav, A.K.; Shukla, R.; Singh, T.R. Machine Learning, Big Data, and IoT for Medical Informatics. In *Chapter 11 Machine Learning in Expert Systems for Disease Diagnostics in Human Healthcare*; Kumar, P., Kumar, Y., Tawhid, M.A., Eds.; Academic Press: Cambridge, MA, USA, 2021; pp. 179–200. ISBN 9780128217771. [CrossRef]
6. Robbins, T.D.; Lim Choi Keung, S.N.; Sankar, S.; Randeva, H.; Arvanitis, T.N. Risk factors for readmission of inpatients with diabetes: A systematic review. *J. Diabetes Complicat.* **2019**, *33*, 398–405. [CrossRef]
7. Soh, J.G.S.; Wong, W.P.; Mukhopadhyay, A.; Quek, S.C.; Tai, B.C. Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: A systematic review with meta-analysis. *BMJ Open Diabetes Res. Care* **2020**, *8*, e001227. [CrossRef]
8. Rubin, D.J. Correction to: Hospital Readmission of Patients with Diabetes. *Curr. Diabetes Rep.* **2018**, *18*, 21. [CrossRef]
9. Healy, S.J.; Black, D.; Harris, C.; Lorenz, A.; Dungan, K.M. Inpatient diabetes education is associated with less frequent hospital readmission among patients with poor glycemic control. *Diabetes Care* **2013**, *36*, 2960–2967. [CrossRef]
10. Hsu, J.-C.; Wu, F.-H.; Lin, H.-H.; Lee, D.-J.; Chen, Y.-F.; Lin, C.-S. AI Models for Predicting Readmission of Pneumonia Patients within 30 Days after Discharge. *Electronics* **2022**, *11*, 673. [CrossRef]
11. Cui, S.; Wang, D.; Wang, Y.; Yu, P.-W.; Jin, Y. An improved support vector machine-based diabetic readmission prediction. *Comput. Methods Programs Biomed.* **2018**, *166*, 123–135. [CrossRef]
12. Dafrallah, S.; Akhloufi, M.A. Factors Associated with Unplanned Hospital Readmission after Discharge: A Descriptive and Predictive Study Using Electronic Health Record Data. *BioMedInformatics* **2024**, *4*, 219–235. [CrossRef]
13. Pham, M.-K.; Mai, T.T.; Crane, M.; Ebiele, M.; Brennan, R.; Ward, M.E.; Geary, U.; McDonald, N.; Bezbradica, M. Forecasting Patient Early Readmission from Irish Hospital Discharge Records Using Conventional Machine Learning Models. *Diagnostics* **2024**, *14*, 2405. [CrossRef] [PubMed]
14. Li, J.; Jiang, X.; Wang, K. Dynamic Partitioning of Graphs Based on Multivariate Blood Glucose Data—A Graph Neural Network Model for Diabetes Prediction. *Electronics* **2024**, *13*, 3727. [CrossRef]
15. Zhu, T.; Kuang, L.; Piao, C.; Zeng, J.; Li, K.; Georgiou, P. Population-Specific Glucose Prediction in Diabetes Care With Transformer-Based Deep Learning on the Edge. *IEEE Trans. Biomed. Circuits Syst.* **2024**, *18*, 236–246. [CrossRef] [PubMed]
16. Nazih, W.; Aseeri, A.O.; Atallah, O.Y.; El-Sappagh, S. Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images. *IEEE Access* **2023**, *11*, 117546–117561. [CrossRef]
17. Chiu, C.-C.; Wu, C.-M.; Chien, T.-N.; Kao, L.-J.; Li, C. Predicting ICU Readmission from Electronic Health Records via BERTopic with Long Short Term Memory Network Approach. *J. Clin. Med.* **2024**, *13*, 5503. [CrossRef]
18. Lu, H.; Uddin, S. Explainable Stacking-Based Model for Predicting Hospital Readmission for Diabetic Patients. *Information* **2022**, *13*, 436. [CrossRef]
19. Strack, B.; DeShazo, J.P.; Gennings, C.; Olmo, J.L.; Ventura, S.; Cios, K.J.; Clore, J.N. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *Biomed Res. Int.* **2014**, *2014*, 781670. [CrossRef]
20. Artetxe, A.; Beristain, A.; Graña, M. Predictive models for hospital readmission risk: A systematic review of methods. *Comput. Methods Programs Biomed.* **2018**, *164*, 49–64. [CrossRef]
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30, pp. 6000–6010.
22. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A.M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the ACL 2017, System Demonstrations*; Association for Computational Linguistics: Vancouver, QC, Canada, 2017; pp. 67–72. [CrossRef]
23. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
24. Louppe, G.; Wehenkel, L.; Sutera, A.; Geurts, P. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*; Burges, C.J., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates, Inc.: New York, NY, USA, 2013; p. 26.
25. Wexler, D.J. Initial Management of Hyperglycemia in Adults with Type 2 Diabetes Mellitus. *UpToDate*. March 2024. Available online: https://www.uptodate.com/contents/initial-management-of-hyperglycemia-in-adults-with-type-2-diabetes-mellitus (accessed on 17 December 2024).
26. Wexler, D.J. Management of Persistent Hyperglycemia in Type 2 Diabetes Mellitus. *UpToDate*. May 2024. Available online: https://www.uptodate.com/contents/management-of-persistent-hyperglycemia-in-type-2-diabetes-mellitus (accessed on 17 December 2024).
27. Dong, T.; Cursio, J.F.; Qadir, S.; Lindenauer, p.K.; Ruhnke, G.W. Discharge disposition as an independent predictor of readmission among patients hospitalised for community-acquired pneumonia. *Int. J. Clin. Pract.* **2017**, *71*, e12935. [CrossRef]
28. Henke, R.M.; Karaca, Z.; Jackson, P.; Marder, W.D.; Wong, H.S. Discharge Planning and Hospital Readmissions. *Med. Care Res. Rev.* **2017**, *74*, 345–368. [CrossRef]

29. Chin, D.L.; Bang, H.; Manickam, R.N.; Romano, P.S. Rethinking Thirty-Day Hospital Readmissions: Shorter Intervals Might Be Better Indicators of Quality of Care. *Health Aff.* **2016**, *35*, 1867–1875. [CrossRef] [PubMed]
30. American Diabetes Association. 15. Diabetes Care in the Hospital: Standards of Medical Care in Diabetes—2021. *Diabetes Care* **2021**, *44* (Suppl. 1), 211–220. [CrossRef] [PubMed]
31. Umpierrez, G.E.; Reyes, D.; Smiley, D.; Hermayer, K.; Khan, A.; Olson, D.E.; Pasquel, F.; Jacobs, S.; Newton, C.; Peng, L.; et al. Hospital Discharge Algorithm Based on Admission HbA1c for the Management of Patients with Type 2 Diabetes. *Diabetes Care* **2014**, *37*, 2934–2939. [CrossRef] [PubMed]
32. Arnold, P.; Scheurer, D.; Dake, A.W.; Hedgpeth, A.; Hutto, A.; Colquitt, C.; Hermayer, K.L. Hospital Guidelines for Diabetes Management and the Joint Commission-American Diabetes Association Inpatient Diabetes Certification. *Am. J. Med. Sci.* **2016**, *351*, 333–341. [CrossRef] [PubMed]
33. Shacham, E.C.; Nitzan, R.; Schwartz, N.; Ishay, A. Effects of Recommendations for Diabetes Management at Hospital Discharge on Long-Term Diabetes Control. *Endocr. Pract. Off. J. Am. Coll. Endocrinol. Am. Assoc. Clin. Endocrinol.* **2021**, *27*, 118–123. [CrossRef]
34. Rubin, D.J.; Maliakkal, N.; Zhao, H.; Miller, E.E. Hospital Readmission Risk and Risk Factors of People with a Primary or Secondary Discharge Diagnosis of Diabetes. *J. Clin. Med.* **2023**, *12*, 1274. [CrossRef]
35. Cuervo Pinto, R.; Álvarez-Rodríguez, E.; González Pérez de Villar, N.; Artola-Menéndez, S.; Girbés Borrás, J.; Mata-Cases, M.; Galindo Rubio, M.; Puig Larrosa, J.; Muñoz Albert, R.; Díaz Pérez, J.A. Documento de consenso sobre el manejo al alta desde urgencias del paciente diabético (Consensus document about discharge management of the diabetic patient from emegency department). *Emergencias* **2017**, *29*, 343–351.
36. Demidowich, A.P.; Batty, K.; Zilbermint, M. Instituting a Successful Discharge Plan for Patients with Type 2 Diabetes: Challenges and Solutions. *Diabetes Spectr.* **2022**, *35*, 440–451. [CrossRef]