POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Cross-domain Textual Explanation for Explainable AI

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Author: **Athira Selvan**

Student ID: 938629
Advisor: Prof. Mark James Carman
Academic Year: 2022-23

Abstract

From the development of self-driving cars to smart assistants, Artificial Intelligence has become a part of most systems we use in our everyday life. Machine learning algorithms used in these systems are black-box models, whose internal working is unknown. Explaining or interpreting the outputs of these models is not possible. Many experts remain wary of using machine learning due to this concern, especially in the domains where these predictions are crucial for decision-making. This makes explainable AI an important field as it provides tools and methods to explain these models. Many works have been done in xAI to produce explanations. However, for normal users who are not from scientific domains, understanding these visual explanations would be another hurdle. So producing natural language explanations for the machine learning models is important. In this work, we develop a system that produces textual explanations for the classification problem. A grammar that was already developed as part of the initial research on this topic is used to generate new training datasets from new domains. In the previous work, GPT-2 model was fine-tuned on cardiovascular and diabetes datasets to produce the textual explanation. Even though the results were promising for explaining datasets from the said domains, the models failed to generalize for new domains. In this work, we further develop the system to make it more generalized to produce textual explanations for classification models from any domain. We add 6 new datasets from multiple domains for training the models. We introduce a modified encoding for the inputs and modified grammar for developing outputs for training. We experimented with the T5 language model for text generation. The Results of the comparative study done on GPT-2 and T5 models show that the T5 model is best suited for this task. We present a multi-domain textual explanation model fine-tuned on T5 that can produce textual explanations for classification models from any domain. We also explore ways to make the model produce more meaningful and varying natural language outputs different from the grammar.

Keywords: explainable AI, textual explanations, xAI

Abstract in lingua italiana

Dallo sviluppo di auto a guida autonoma agli assistenti intelligenti, l'intelligenza artificiale è diventata una parte di molti dei sistemi che usiamo nella nostra vita quotidiana. Gli algoritmi di apprendimento automatico utilizzati in questi sistemi sono modelli black-box, il cui funzionamento interno è sconosciuto. Non è possibile chiarire o interpretare gli output di questi modelli. Molti esperti rimangono cauti nei confronti dell'utilizzo dell'apprendimento automatico a causa di questa preoccupazione, specialmente nei domini in cui queste previsioni sono cruciali per il processo decisionale. Questo rende l'intelligenza artificiale spiegabile un campo importante in quanto fornisce strumenti e metodi per spiegare questi modelli. Molti lavori sono stati fatti nell'xAI per produrre spiegazioni. Tuttavia, per gli utenti normali che non hanno conoscenze in campi scientifici, comprendere queste spiegazioni visive sarebbe un altro ostacolo. Quindi produrre spiegazioni in linguaggio naturale, per i modelli di apprendimento automatico, è importante. In questo lavoro, sviluppiamo un sistema che produce spiegazioni testuali per il problema di classificazione. Una grammatica, che era stata già sviluppata come parte della ricerca iniziale su questo argomento, è usata per generare nuovi insiemi di dati di addestramento da nuovi domini. Nel lavoro precedente, il modello GPT-2 è stato ottimizzato su insiemi di dati cardiovascolari e diabetici per produrre la spiegazione testuale. Anche se i risultati sono stati promettenti per spiegare gli insiemi di dati dei suddetti domini, i modelli non sono riusciti a generalizzare per nuovi domini. In questo lavoro, sviluppiamo ulteriormente il sistema per renderlo più generalizzato, per produrre spiegazioni testuali per i modelli di classificazione da qualsiasi dominio. Aggiungiamo 6 nuovi insiemi di dati da diversi domini per addestrare i modelli. Introduciamo una codifica modificata per gli input e una grammatica modificata per sviluppare outputs per l'addestramento. Abbiamo sperimentato il modello di linguaggio T5 per la generazione del testo. I risultati dello studio comparativo condotto sui modelli GPT-2 e T5 mostrano che il modello T5 è più adatto per questo compito. Presentiamo un modello di spiegazione testuale multidominio ottimizzato su T5 che può produrre spiegazioni testuali per i modelli di classificazione da qualsiasi dominio. Esploriamo anche modi per far produrre al modello outputs diversi, dalla grammatica, in linguaggio naturale, più significativi e vari.

‘**Parole chiave:** AI spiegabile, spiegazioni testuali, xAI

Contents

1 | Introduction

As the field of artificial intelligence gets more and more popular, we find it being used in almost all technologies and domains these days. From voice assistants to self-driving cars, we see AI algorithms in every technology we use nowadays. Most of these systems are black-box models. The exact justifications for why these models produced a particular output are not known. A lot of work is being done for making these models more justifiable, as that information is very crucial when using these models in important decision-making tasks. But for that to be done, these models need to be more simple, which in turn affects their performance. This is a reason why machine learning models are not being widely used in the field of medicine, security, etc. When making a diagnosis decision, the doctor needs to know what factors made the model make that decision. This information is very important as a wrong diagnosis can even threaten the patient's life. So explainable AI, a field that deals with methods and tools to explain the black-box models, has become more popular and important for making machine learning useful in real-life. xAI has many existing tools for explaining these models. All the explanations given by these tools are numerical or graphical. For common users, the most preferred mode to receive an explanation is through natural language. A patient will prefer to get a detailed report in natural language about his diagnosis rather than in tables or graphical forms.

2 | Background & Related Works

2.1. Artificial Intelligence and Machine Learning

2.1.1. Machine Learning

From the development of self-driving cars to smart assistants, Artificial Intelligence has become an important part of our everyday life. Machine learning is a subdomain of Artificial Intelligence. Machine Learning Proposes techniques and frameworks to extract knowledge from data. It focuses on Data-driven techniques, in opposition to model- or expert-driven techniques. There are three learning paradigms:

- Supervised Learning: Supervised learning is the most widely used sub-field of machine learning. Given training data set $D = \{x_i, t_i\}_{i=1,...,N}$ including desired outputs from some unknown function f. The aim is to find a good approximation of f that generalizes well on test data. Input variables x_i are called features or attributes. Output variables t_i are called targets or labels. If t_i is discrete the task is called classification. if t_i is continuous it is called regression. if t_i is the probability of x_i it is called probability estimation, which is different from regression because of the constraints imposed by probability.
- Unsupervised Learning: Given a set of inputs, learn to exploit regularities in the inputs to build a new representation of them.
- Reinforcement Learning: Given a set of actions to perform in an environment and the corresponding rewards, learn how to maximize cumulative reward through the actions.

We focus on a supervised learning approach for our task. In particular, we focus on the classification task. There are many algorithms in machine learning for classification.

The goal of classification is to assign an input x into one of K discrete classes C_k, where k = 1, ..., K. we use 3 classification models in this thesis.

- XGBoost

- Random Forest
- Logistic Regression

2.2. Black-box AI

Generally, the Black Box Problem can be defined as an inability to fully understand an AI's decision-making process and the inability to predict the AI's decisions or outputs.[5] Many machine learning (ML) algorithms used to develop the systems are inscrutable, particularly deep learning neural network approaches which have emerged to be a very popular class of ML algorithms. This inscrutability can hamper users' trust in the system, especially in contexts where the consequences are significant and lead to the rejection of the systems. [17] There exist a performance and explainability tradeoff paradigm in machine learning. When the performance of the model is high the interpretability or explainability of these models becomes low. The black-box models like deep learning and ensembles show high performance but their transparency is very low making users blind about how these decisions were taken. And then there we have the white-box or glass-box models like linear and decision-tree-based models, whose predictions are explainable with common examples. But they fail to achieve state-of-the-art performance when compared to the high-performing black-box models. Many remain wary of machine learning due to this concern. We need a model that is both trustworthy and high performing to be used in real-life scenarios. This led to the emergence of eXplainable Artificial Intelligence. Explainable AI popularly known as xAI is a field that deals with developing tools and methods to explain the decisions of these black-box models.

2.3. eXplainable AI

A lot of research is being done now to tackle the explainability issue of machine learning. xAI had been unpopular in the past as more focus was given to improving the predictive capability of the AI models. The need of understanding "why" for the predictions made by the model was not given much attention. But when these models became popular and performed well enough to be used in real-life scenarios, it gave rise to the question about the trustability of the model decisions. Especially in the fields of medicine and security the experts were critical of using the services of these models. And this gave rise to the popularity of xAI.

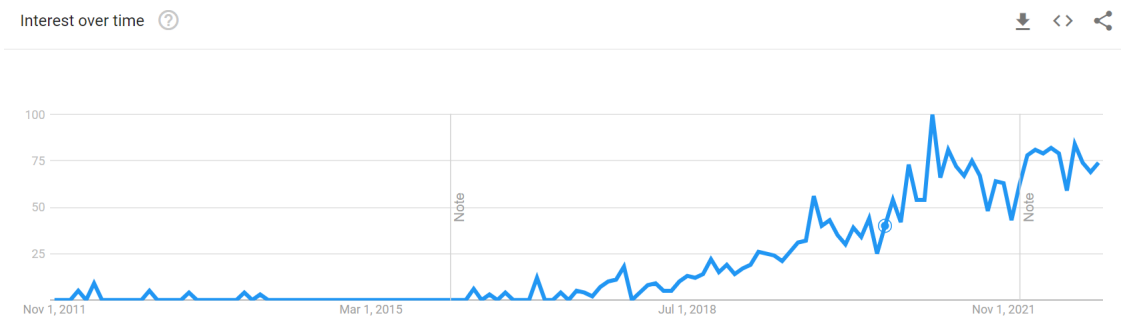

Figure 2.1: google trend of "explainable AI" from 2011 to 2022

Many approaches have been proposed for interpreting the black-box models. Some methods are mentioned below.

- **SHAP**

  Shapley Additive Explanations[10] is based on the Shapley value concept of game theory. It is based on the concept of the importance of each feature to the overall coalitions.

- **LIME**

  Local Interpretable Model-agnostic Explanations are based on the principles of sampling and obtaining a surrogate dataset and then selecting features from the surrogate dataset. It selects top features based on techniques like Lasso. .

- **Counterfactual explanation**

  Counterfactual explanation explains what minimum changes in the features can cause a change in the prediction. It says that if the following feature values were changed to these values then the prediction would have been this. This information is taken from the dataset by finding the closest data point that belongs to another class.

- **Ceteris paribus**

  Ceteris Paribus profiles show how changing the value of one feature, keeps all others as constant changes in the output of the model.

2.4. Textual Explanations

Many works have been done in xAI to produce visual explanations. However, for normal users who are not from scientific domains, understanding these visual explanations would be another hurdle. So it is important to have these explanations in a more understandable form of natural language text. So producing natural language explanations for the machine learning models is important. This is a task of natural language generation and comes under the field of natural language processing.

2.4.1. Natural Language Generation

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze amounts of natural language data. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

Natural language understanding is a subset of natural language processing. Syntax and semantic analysis is done on text and speech to understand the meaning of the texts. Natural language generation is another subset of natural language processing that generates the human language text from structured data.

2.4.2. Hallucination in NLG

Hallucination in natural language generation happens when the model-generated output contains text that is not related to the input. There are 2 types of hallucination for our task - Intrinsic and Extrinsic hallucinations. An intrinsic hallucination, which is verifiable, happens when the generated output contains information that is contradicted by the input data. Extrinsic hallucinations, which cannot be verified, happen when the generated text contains extra information which is not related to the input. [9] Hallucination in natural language generation is a well-known problem. As we need our model to generate explanations for new unseen domains, we are faced with this challenge. We want the model to be more generalized but at the same time, we need it to learn to use correct values and feature names given in the input. We are okay with the model modeling the feature name to something with the same meaning, but the model completely inventing new features and values is not acceptable while explaining.

An example of intrinsic hallucination in our case is when the model says systolic blood

pressure is 130, even though it was 90 in the input. An example of extrinsic hallucination is when it uses phrases like "We are all about making sure we get enough information for this prediction", this phrase was invented by the model which was not in our grammar. So we do not have a way to evaluate this kind of phrase using any metrics.

2.5. Language Models

A transformer is a model that is based only on the attention mechanism and is composed of a stack of encoder/decoder layers. A language model is a probabilistic model that is capable of predicting the next word in a sentence given the word (s) preceding it. The idea was to use a deep autoencoder to build a non-linear and continuous language model. These models can be used for verities of tasks like translation, question answering, etc. We use 2 well-known language models GPT2 and T5 for text generation.

2.5.1. GPT2

Generative Pre-trained Transformer 2 (GPT-2) [15] transformer-based language model by OpenAI. It is trained to be used for translation, question answering, summarization, text generation, etc. The model is trained on various domains making it efficient for text generation in various domains. GPT-2 has way more parameters and data than its predecessor GPT.

2.5.2. T5

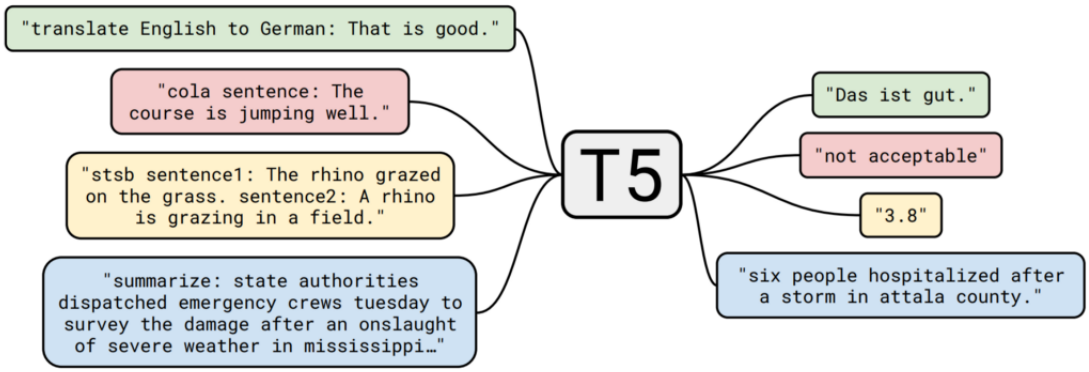

Figure 2.2: T5 model [16]

T5 is an encoder-decoder model. It is pre-trained in text-to-text format. T5 can be finetuned for a variety of tasks by adding a different prefix to the input of each task. T5 can be readily used for pre-trained tasks like translation, summarization, etc.

2.6. Related works on textual explanation

In the following section, we detail the related works done on textual explanation.

TextVQA

Nagaraj-rao-etal[14] proposes MTXNet, an end-to-end trainable multimodal architecture. The generated explanations focus on the text in the image. The work focuses on giving explanations only for images. In our work, we deal with classification datasets that are tabular. Our model gives explanations based on all features and the other data instances present in the dataset. We make use of the xAI techniques of SHAP, counterfactuals, etc for the explanation.

The natural language explanation algorithms for the lung cancer computer-aided diagnosis system

The paper [13] proposes 2 algorithms for natural explanations for decisions of a lung cancer computer-aided diagnosis system. The first part of their algorithm uses LIME for selecting important features and the second part converts the important features to natural explanations. The first algorithm uses a special vocabulary of simple phrases which produce sentences and their embeddings. The second algorithm reduces the problem to a set of simple classifiers. The work only focuses on lung scan images and the explanation is generated from the algorithm. In our thesis, we focus on classification datasets from many different domains and we make use of a language model in order to generate the textual explanation.

3 | Research questions

The purpose of this work is to answer the following research questions:

- **Do introducing training data from different domains allow the models to generalize better?**

  Training the GPT2 on the cardiovascular dataset was proved to produce a textual explanation for inputs given from the same domain. But it was not able to give good results for the diabetes dataset. The research focused on making the model produce a textual explanation for any given dataset.

  We also need to answer these sub-questions in order to make this model more multi-domain.

  a) How many new datasets should be introduced?

  As more training data will surely make the model perform better, the question arises of how many is enough for making the model start giving good outputs for a new unseen dataset.

  b) More small datasets or more large datasets with many instances?

  We need to understand the balance between more variety or more training sets from the same domains, in order to reduce the over-fitting of the model.

- **Which transformer architecture is most appropriate for the explanation generation task in terms of performance?**

  There are different transformer architectures like encoder-only, encoder-decoder, or decoder-only. There are different language models available that use different architectures. Popular ones are GPT2, T5, Bert, etc. GPT2 was used as the pre-trained model in the previous work. We need to explore how other models perform in this task in order to choose the best one.

- **In other words, is the explanation task best modeled as a text generation or text translation task?**

Another interesting aspect of this task is understanding the subdomain of natural language generation it belongs to. On one hand, transforming the input given in the encoded form into a meaningful explanation in natural language can be considered a translation task. On the other hand, the same can be counted as a text generation task as we give the model more freedom to include sentences in the explanation that do not necessarily have to be in the input. More information regarding the disease or a particular feature that can help justify the prediction is always preferred.

- **How can we evaluate explanations in natural language?**

  Evaluating the quality of output of the text generation models is a research topic in itself. Even though there are many evaluation metrics like BLEU, METEOR, etc, we cannot rely on these metrics alone to evaluate sentences that can vary from the references. In our case, we want the model to produce more natural outputs different from those produced by the grammar. That is also the prime reason why we use a Language model rather than using grammar alone for the explanations. So we need new metrics which can accurately measure the hallucination in the produced output. We need the explanation to be accurate in terms of the values of the feature given in the input.

- **How important is language model pre-training?**

  As we fine-tune the language model with our explanation dataset generated by the grammar, a question arises about the importance of this training step. Because these models, given the amount of data it is trained on, have excellent text-generation capabilities, We need to explore the methods to make the models generate explanations different from the ones given during the training. We want these models to produce more natural explanations different from the grammar. So we need to access how the training affects the generalization capabilities of the language models.

- **How to diversify the generated explanations?**

  We need explanations to be more diverse from the explanations generated from the grammar. This is also another reason why we choose to use language models to generate explanations. We need to explore ways to make the generated outputs more diverse.

- **Hallucination-generalization trade-off and How to evaluate and mitigate the hallucination in the generated output?**

  Evaluating hallucination in the generated output is a hard task. As we expect the model to produce a more generalized output, we need it to be very different from the

reference output we generate using grammar. But on the other hand, it will mean that we need the model to generate output different from references but we also need the model to be consistent with the feature values given in the output. This is a trade-off between both. We need to explore more ways to mitigate hallucination in such a way that it does not affect the generalization capability of the model.

4 | Datasets

In this chapter, we give details on all the datasets used for making the training set. Datasets that provide textual explanations for machine learning model decisions are not readily available. We need a dataset that has explanations for a decision. The only dataset that looks close to what we need is MedDialog. Even though MedDialog [20] provides a dataset with patient-doctor conversations, extracting diagnosis explanation from that was not possible. Therefore to train our model we need to create explanations using grammar. As we are focusing on cross-domain training, we need classification datasets from different domains. Below we list all the datasets we used for this thesis.

4.1. Cardiovascular

Cardiovascular diseases are a group of disorders of the heart and blood vessels. Many factors affect cardiovascular health, mainly, diet, physical activity, smoking, alcohol consumption, etc. There are also some objective features that affect cardiovascular health, such as age, Body Mass Index, raised blood pressure, raised blood glucose, etc.

When a patient is diagnosed as positive for cardiovascular disease, it is very useful for the patient to understand what factors led o this prediction and what can be done to reduce the risk. This information will help the patient to take informed decisions toward a healthy lifestyle. For example, obesity can cause heart disease. It is useful for the patient to know that his Body Mass Index indicates that he is obese and that reducing it to a normal level can reduce his risks.

For training, we use a dataset taken from Kaggle [1] with 11 features and 1 target. It is a huge dataset with data from 70,000 patients collected at the moment of medical examination.

3 types of features are available.

- Objective Feature: Information about the patient.

  Age, Height, Weight, and Gender.

- Examination Feature: Taken from the medical examination report of the patient.

  Systolic blood pressure, Diastolic blood pressure, Cholesterol, and Glucose. The first two are numerical and the last 2 are categorical features.

- Subjective Feature: Given by the patient.

  smoking, alcohol, Physical activity, and cardiovascular disease. All of them are binary features. The last one is the target value in our dataset.

For the experiments, the dataset was cleaned. The height and weight values were combined into a single feature named BMI and the outliers were removed from the dataset.

	age	gender	systolic blood pressure	diastolic blood pressure	cholesterol	glucose	smoking	alcohol	physical activity	cardio	BMI
0	50.391781	2	110	80	1	1	0	0	1	0	21.967120
1	55.419178	1	140	90	3	1	0	0	1	1	34.927679
2	51.663014	1	130	70	3	1	0	0	0	1	23.507805
3	48.282192	2	150	100	1	1	0	0	1	1	28.710479
4	47.873973	1	100	60	1	1	0	0	0	0	23.011177

Figure 4.1: Cardio dataset sample

Cholesterol and glucose values are in the range of 1-3, indicating Normal, Above Normal, and Well Above Normal levels respectively. cardio value 1 indicates that cardiovascular disease is present. Value 0 indicates the absence of disease. For smoking, alcohol, and physical activity features, 0 indicates yes and 1 indicates no. 1 indicates the gender is female and 2 indicates it is male.

Even though the dataset has 70000 values only 7000 balanced instances were used for training the model, to avoid overfitting as all the other datasets had less than 10000 instances.

4.2. Stroke

Stroke is the second leading cause of global death according to WHO.

The dataset [8] is taken from Kaggle and is used to predict whether a person is likely to get a stroke. The features are the following:

1. Gender: It is a categorical feature with values "Male", "Female" or "Other"
2. Age: Numerical features with age in years
3. hypertension: Binary feature indicated by 0 for no, 1 for yes

4. heart disease: Binary feature indicated by 0 for no, 1 for yes
5. ever married: categorical features with values "No" or "Yes"
6. work type: Type of work done "children", "Government job", "Never worked", "Private" or "Self-employed"
7. Residence type: Area where the person lives - "Rural" or "Urban"
8. average glucose level: average glucose level in blood
9. BMI: Body Mass Index
10. smoking status: smoking status of the person "formerly smoked", "never smoked", "smokes" or "Unknown"
11. stroke: Target variable which takes value 1 if the patient is likely to get a stroke or 0 if not

The dataset contains 5110 instances. All instances with null and unknown are removed before training. Sample data:

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked
Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked
Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked
Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes
Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked

Figure 4.2: Stroke dataset sample

4.3. Breast Cancer

A digitalized image of a Fine Needle Aspirate of a breast mass is used to compute the features of this dataset. Breast cancer is a form of cancer occurring in the cells of the breasts. If a suspicious lump is found during self-examination or an x-ray, a diagnosis is conducted to determine if is cancerous or not. This breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison by Dr. William H. Wolberg. [12]

This dataset contains the following features:

1. mean_radius
2. mean_texture

3. mean_perimeter
4. mean_area
5. mean_smoothness
6. diagnosis

Diagnosis is our target variable. All 5 features are numerical features giving information about the lump. Diagnosis can be benign or malignant denoting breast cancer or no breast cancer respectfully. 0 indicates benign and 1 indicates malignant. It has 569 instances with 212 benign and 356 malignant classes.

A sample dataset is given below:

mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
17.99	10.38	122.8	1001	0.1184	0
20.57	17.77	132.9	1326	0.08474	0
19.69	21.25	130	1203	0.1096	0
11.42	20.38	77.58	386.1	0.1425	0
20.29	14.34	135.1	1297	0.1003	0

Figure 4.3: Breast cancer dataset sample

4.4. Mammographic mass

Mammography is a breast cancer screening method. The dataset [11] can be used to predict the severity (benign or malignant) of a mammographic mass. The BI-RADS attributes and the patient's age can be used for this prediction as features. The dataset contains a BI-RADS assessment, the patient's age, three BI-RADS attributes, and the diagnosis. The dataset has 516 benign and 445 malignant masses.

BI-RADS	Age	Shape	Margin	Density	Severity
5	67	3	5	3	1
5	58	4	5	3	1
4	28	1	1	3	0
5	57	1	5	3	1
5	76	1	4	3	1
3	42	2	1	3	1

Figure 4.4: mammographic mass dataset sample

4.5. Statlog heart disease

This dataset taken from [7] has many features. We use only 10 important features in order to reduce the number of features. We have:

1. age
2. sex - male/female
3. chest pain type
4. maximum heart rate
5. exercise-induced angina
6. old peak
7. slope
8. number of colored vessels
9. thallium
10. presence

The feature prediction is our target value. There are 270 instances with 150 positive and 120 negative predictions.

age	sex	type chest pain	maximum heart rate	exercise induced an	oldpeak	slope	number of colored v	thallium	presence
70	1	4	109	0	2.4	2	3	3	0
67	0	3	160	0	1.6	2	0	7	1
57	1	2	141	0	0.3	1	0	7	0
64	1	4	105	1	0.2	2	1	7	1
74	0	2	121	1	0.2	1	1	3	1
65	1	4	140	0	0.4	1	0	7	1

Figure 4.5: heart disease dataset sample

4.6. Occupancy detection

This dataset [3] is used for predicting room occupancy based on the internet of things sensor data. Understanding if the room is occupied or not is very useful for automatic electronic device functions. This Dataset taken from kaggle contains 2666 rows and 6 columns with features: Temperature, Humidity, Light, CO2, Humidity ratio - all 5 numerical features and the target variable Occupancy.

ID	date	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
1	04-02-2015 17:51	23.18	27.272	426	721.25	0.004792988	1
2	04-02-2015 17:51	23.15	27.2675	429.5	714	0.004783441	1
3	04-02-2015 17:53	23.15	27.245	426	713.5	0.004779464	1
4	04-02-2015 17:54	23.15	27.2	426	708.25	0.004771509	1
5	04-02-2015 17:55	23.1	27.2	426	704.5	0.004756993	1
6	04-02-2015 17:55	23.1	27.2	419	701	0.004756993	1
7	04-02-2015 17:57	23.1	27.2	419	701.6666667	0.004756993	1
8	04-02-2015 17:57	23.1	27.2	419	699	0.004756993	1

Figure 4.6: occupancy dataset sample

4.7. Diabetes

Diabetes is a chronic health condition. It affects the process of metabolism and causes high sugar levels in the blood.

The dataset taken from kaggle [2] consist of several medical predictor variables like:

1. number of pregnancies
2. BMI
3. insulin level
4. age
5. blood pressure
6. skin thickness
7. glucose
8. Diabetes pedigree function
9. Outcome

The outcome is our target variable. 268 of 768 instances are positive and the others are negative

Age	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree F	Outcome
50	6	148	72	35	125	33.6	0.627	1
31	1	85	66	29	125	26.6	0.351	0
32	8	183	64	29	125	23.3	0.672	1
21	1	89	66	23	94	28.1	0.167	0
33	0	137	40	35	168	43.1	2.288	1
30	5	116	74	29	125	25.6	0.201	0
26	3	78	50	32	88	31	0.248	1

Figure 4.7: Diabetes dataset sample

4.8. Smoke detection

This dataset [4] taken from Kaggle is used to predict if a fire alarm will go on or not based on the values of the features. The features are values coming from different IoT sensors measuring different conditions of the room to determine if there is smoke. This dataset has about 60000 rows and 16 columns. We were not able to use this dataset for training due to its long input encoding because of the high number of features. But w used it to test the generalization capability of the model on the new dataset.

UTC	Temperature[C]	Humidity[%]	TVOC[ppb]	eCO2[ppm]	Raw H2	Raw Ethanol	Pressure[hPa]	PM1.0	PM2.5	NC0.5	NC1.0
1654733331	20	57.36	0	400	12306	18520	939.735	0	0	0	0
1654733332	20.015	56.67	0	400	12345	18651	939.744	0	0	0	0
1654733333	20.029	55.96	0	400	12374	18764	939.738	0	0	0	0
1654733334	20.044	55.28	0	400	12390	18849	939.736	0	0	0	0
1654733335	20.059	54.69	0	400	12403	18921	939.744	0	0	0	0
1654733336	20.073	54.12	0	400	12419	18998	939.725	0	0	0	0
1654733337	20.088	53.61	0	400	12432	19058	939.738	0	0	0	0
1654733338	20.103	53.2	0	400	12439	19114	939.758	0	0	0	0
1654733339	20.117	52.81	0	400	12448	19155	939.758	0	0	0	0
1654733340	20.132	52.46	0	400	12453	19195	939.756	0.9	3.78	0	4.369

Figure 4.8: Smoke dataset sample

5 | Approach

This chapter details our approach to cross-domain textual explanation.

5.1. Previous work

The work in this thesis is a continuation of Vittorio Torri's Master's thesis "Textual explanation for Intuitive Machine Learning" [19]. The previous work developed a textual explanation model using GPT2. As there was no training set already available, the thesis developed a grammar to create the training dataset. The grammar was made for the cardiovascular dataset.

Below we list the workflow of the previous work.

1. 4 classification models were trained using the cardio dataset, in order to predict if a patient has cardiovascular disease or not.
2. xAI methods like SHAP values, counterfactuals, and ceteris paribus were considered for generating the explanations.
3. The input feature values, along with SHAP values of the features, mean and standard deviation values for numerical features, and one counterfactual feature were the inputs given to the model in order to produce the explanation.
4. The input was encoded in a format. given below
5. A grammar was prepared for generating textual explanations for a given input.
6. The encoded input along with the explanation generated by the grammar as the desired output was contexted together to be trained with GPT2.
7. A part of this dataset was kept aside for testing.
8. After training GPT2 with this dataset, the model was tested using BLEU, METEOR, BLEURT scores for the test set from the kept-aside training set.
9. Model was also tested on a new dataset of Pima diabetes. For testing, this dataset

was also encoded and a new grammar was defined for the diabetes dataset for providing references.

The results from the work showed that the model performed well for the cardiovascular dataset but failed to generalize for the diabetes dataset.

To handle that, the model was fine-tuned on the diabetes dataset. The results showed that after fine-tuning the model performed well for diabetes data.

In this thesis we make the model generalize such that it produces textual explanations for given any dataset from any domain. A detailed explanation of our experiments and progress is given in the next sections.

5.2. Multi domain training

The first step focused on training the model with more datasets. A modified grammar was written for each new dataset. Six new datasets were added for the research. Initially, the aim was to see the effects on the performance of the model when trained on 2 datasets from different domains and tested on a dataset from a new domain. We started out with cardio, stroke, and diabetes datasets. The following sets of experiments were done using the GPT2 model. Only a few instances of the cardio dataset were used in order to limit overfitting.

- Cardio + Stroke

  We first trained the model on about 5000 instances of stroke and about 7000 instances of cardio and was tested on the diabetes dataset to see if it is able to generalize.

- stroke + pima

  The model was then trained on almost 5000 instances of stroke and almost 700 instances from the diabetes dataset. The model was trained for multiple epochs and was tested on Pima, cardio, and stroke datasets.

- cardio + pima

  We also tried to train it with both Pima and cardio, using only 7000 instances from cardio for training.

5.3. New Grammars

As there was no training dataset readily available for this task, we had to construct grammars that produces explanations for a model decision. We give the grammars the input features, model output, SHAP values, and counterfactuals. The grammar produces explanations based on these inputs.

Six new datasets were introduced for training and testing purposes. A new grammar was introduced for each of them to create the training and test sets.

In this section, we present 3 grammar used for making training sets for 3 domains. You can find the other three grammar in Appendix A. All these grammars were developed by making modifications to the grammar presented in [19]

5.3.1. Stroke

- **Grammar**

  The grammar used for the stroke dataset is given below. New rules were written for categorical features heart disease, hypertension, married, work type, and residence.

```
S -> F SN T CF

F ->The main reason why P has been predicted as O1 is the EF |
    The first motivation for the prediction of O2 is the EF |
    The O2 outcome is primarily determined by the EF |
    The EF is a significant factor that determines the outcome of
    O2 |
    The first cause determining the outcome of O2 is the EF |
    Of primary importance for predicting O2 is the EF |
    The first element which influenced the prediction of O2 is the
    EF |
    The EF is important for predicting O2 |
    The diagnosis of O1 for this patient is based on the EF |
    The most relevant factor for the prediction of O2 is the EF

SN -> . In addition , the EF also has a significant influence |
      . The EF is also an important element |
      . Moreover, the EF plays an important role
```

T −> and also the EF is relevant . |
, while the third factor is the EF. |
. Finally , the EF also influences the result .

P −> he | she

O1 −> having low probability of getting a stroke with a confidence of PER% | having a high probability of getting a stroke with a confidence of PER%

O2 −> low probability of getting a stroke with a confidence of PER% | high probability of getting a stroke with a confidence of PER%

EF −> AGE | OTHER-NUM | HD | HT | GENDER | SMOKING | MARRIED | WORKTYPE | RESIDENCE

AGE −> fact that patient is elderly WHY | fact that the patient is young WHY | fact that the patient is middle-aged WHY

HD −> fact that the patient had a heart disease WHY | fact that the patient did not have any heart disease WHY

HT −> fact that the patient had hypertension WHY | fact that the patient did not have hypertension WHY

GENDER −> male gender WHY | female gender WHY | non binary gender WHY

SMOKING −> fact that the patient is a smoker WHY | fact that the patient is not a smoker WHY

MARRIED −> fact that the patient has never married WHY | fact that the patient was married WHY

WORKTYPE −> fact that the patient was a child WHY | fact that the patient was self-employed WHY | fact that the patient worked in a public sector WHY | fact that the patient was unemployed WHY

RESIDENCE −> fact that the patient lived in a rural area
WHY | fact that the patient lived in a
urban area WHY

OTHER-NUM −> value of f (v) DIST WHY

DIST −> , which is n-std standard deviations above the mean, | , which
is n-std standard deviations below the mean, | , which is higher
than the mean, | , which is lower than the mean,

CF −> FIRST-CF-F CF-F

FIRST-CF-F −> If CF-F-DESC

CF-F −> , CF-F-DESC |

CF-F-DESC −> ALCOHOL-CF | SMOKE-CF | GENDER-CF | OTHER-CF |
MARRIED-CF | WORKTYPE-CF | RESIDENCE-CF

ALCOHOL-CF −> the patient drank alcohol | the patient did not drink alcohol

SMOKE-CF −> the patient was a smoker | the patient was not a smoker

GENDER-CF −> the patient was a male | the patient was a female

MARRIED-CF −> the patient has never married | the patient was married

WORKTYPE-CF −> the patient was a child
| the patient was self-employed
| the patient worked in a public sector
| the patient was unemployed

RESIDENCE-CF −> the patient lived in a rural area
|the patient lived in an urban area

OTHER-CF –> f was v

WHY –> , where high values of this feature are associated with a high probability of stroke |, where low values of this feature are associated with a low probability of stroke |, where low values of this feature are associated with a high probability of stroke|, where high values of this feature are associated with a low probability of stroke | ϵ

- **Encoded Input**

The input given to the model is encoded in the following format.

input=[name=gender(Female), shap=-0.0]; [name=age(45), shap=0.0, mean=43.2, std=22.6]; [name=hypertension(no), shap=0.0]; [name=heart disease(no), shap=0.0]; [name=was married(Yes), shap=0.0]; [name=type of work(Govt_job), shap=0.0]; [name=Type of residence(Rural), shap=0.0]; name=average glucose level(68.66), shap=-0.0, mean=106.1, std=45.3]; [name=BMI(25.3), shap=0.0, mean=27.8, std=9.5]; name=smoking(never smoked), shap=0.0];

target=[name=stroke, prediction=low probability of getting stroke, confidence=96%];

cf=[name=smoking(formerly smoked)][name=age(51.0)][name=average glucose level(103.4)];

cf_pred=high probability of getting stroke

- **Output from Grammar**

The first motivation for the prediction of low probability of getting stroke with a confidence of 96% is, the fact that the patient is middle-aged, where high values of this feature are associated with a low probability of getting stroke The second important element is the value of BMI (25.3), where low values of this feature are associated with a low probability of getting stroke and the fact that

the patient did not have hypertension also affects the prediction. If the patient used to smoke, BMI was 27, age was 51 and average glucose level was 103 the result would have been high probability of getting stroke

5.3.2. Breast Cancer

- **Grammar**

  The following is the grammar used to create training outputs for Breast cancer dataset.

```
S -> F SN T CF

F ->The main reason for the prediction O1 is the EF |
    The most important feature which influenced the prediction of
    O2 is the EF |
    The prediction of O2 has high dependence on EF |
    The most relevant factor for the prediction of O2 is the EF |
    The O2 outcome is primarily determined by the EF |
    The EF plays a significant role in O2 |
    Of primary importance for predicting O2 is the EF |
    The EF is important for predicting O2 |
    The EF is most important in predicting O2 |
    The EF is a dominant factor for the prediction of O2

SN -> . In addition, the EF also has a significant influence |
      . The EF is also an important element |
      . Moreover, the EF plays an important role |
      . The second important element is the EF |
      . Another important feature is the EF |
      . Furthermore, the EF has a considerable effect |
      . A second factor to consider is the EF |
      . The EF also has a significant effect |
      . The EF is another major factor
```

T −> and also the EF is relevant. |
, while the third factor is the EF. |
. Finally , the EF also influences the result .|
and the EF also affects the prediction |
and the EF also contributes to the result |
The result is also affected by the EF |
The EF is the third factor that determines the outcome |
It is important to mention the EF as well |
The EF is also worth mentioning among the causes

O1 −> having no breast cancer | having a breast cancer

O2 −> no breast cancer | breast cancer

EF −> OTHER-NUM

OTHER-NUM −> value of f (v) DIST WHY

DIST −> , which is n-std standard deviations above the mean, | , which is n-std standard deviations below the mean, | , which is higher than the mean, | , which is lower than the mean,

CF −> FIRST-CF-F CF-F

FIRST-CF-F −> If CF-F-DESC

CF-F −> , CF-F-DESC |

CF-F-DESC −> OTHER-CF

OTHER-CF −> f was v

WHY −>	The higher the value of EF, lower the chances of detecting breast cancer. | The EF value is quite low which increases the chances of detecting breast cancer. | , because when the EF value increases, there are high chances that mass is malignant | , because low values of EF means that the chances of being malignant is low, | ϵ

- **Encoded Input**

input=[name=mean_radius(12.88), shap=0.0, mean=14.1, std=3.5];[name=mean_texture(18.22), shap=0.1, mean=19.3, std=4.3];[name=mean_perimeter(84.45), shap=0.1, mean=92.0, std=24.3];[name=mean_area(493.1), shap=0.1, mean=654.9, std=351.9];[name=mean_smoothness(0.1218), shap=-0.2, mean=0.1, std=0.0];
target=[name=breast cancer, prediction=breast cancer, confidence=81%];
cf=[name=mean_radius(13.2)]name=mean_texture(18.7)] name=mean_perimeter(86.0)][name=mean_smoothness(0.1)];
cf_pred=negative

- **Output from Grammar**

The most important feature which influenced the prediction of breast cancer with a confidence of 81% is the value of mean perimeter (84), which is lower than the mean value. The mean perimeter value is quite low which increases the chances of detecting breast cancer. In addition, the value of mean texture (18) also has a significant influence. The mean texture value is quite low which increases the chances of detecting breast cancer, while the third factor is the value of mean area (493), which is lower than the mean value. If mean smoothness was 0, mean perimeter was 85 and mean radius was 13 the result would have been no breast cancer.

5.3.3. Smoke

- **Grammar**

The following is the grammar used to create training outputs for stroke dataset.

S −> F SN T CF

F −>The main reason for the prediction O1 is the EF | The most important feature which influenced the prediction of O2 is the EF | The prediction of O2 has high dependence on EF | The most relevant factor for the prediction of O2 is the EF | The O2 outcome is primarily determined by the EF | The EF plays a significant role in O2 | Of primary importance for predicting O2 is the EF | The EF is important for predicting O2 | The EF is most important in predicting O2 | The EF is a dominant factor for the prediction of O2

SN −> . In addition , the EF also has a significant influence | . The EF is also an important element | . Moreover, the EF plays an important role | . The second important element is the EF | . Another important feature is the EF | . Furthermore, the EF has a considerable effect | . A second factor to consider is the EF | . The EF also has a significant effect | . The EF is another major factor

T −> and also the EF is relevant. | , while the third factor is the EF. | . Finally , the EF also influences the result .| and the EF also affects the prediction | and the EF also contributes to the result | The result is also affected by the EF | The EF is third factor that determines the outcome | It is important to mention the EF as well | The EF is also worth mentioning among the causes

O1 −> having no smoke | having a smoke

O2 −> no smoke | smoke

EF −> OTHER-NUM

OTHER-NUM −> value of f (v) DIST WHY

DIST −> , which is n-std standard deviations above the mean, | , which is n-std standard deviations below the mean, | , which is higher than the mean, | , which is lower than the mean,

CF −> FIRST-CF-F CF-F

FIRST-CF-F −> If CF-F-DESC

CF-F −> , CF-F-DESC |

CF-F-DESC −> OTHER-CF

OTHER-CF −> f was v

WHY −> The higher the value of EF, lower the chances of detecting smoke. | The EF value is quite low which increases the chances of detecting smoke. | , because when the EF value increases, there are high chances that there is smoke | , because low values of EF means that the chances of detecting smoke, | ϵ

- **Encoded Input**

input=[name=temperature(28.55), shap=-0.0, mean=16.9, std=15.0];[name=humidity(43.35), shap=0.0, mean=46.8, std=10.3];[name=TVOC(44), shap=0.1, mean=2806.7, std=10523.4];[name=eCO2(417), shap=-0.0, mean=775.9, std=2340.5];[name=Raw H2(12794), shap=0.0, mean=12931.4, std=336.2];[name=Raw Ethanol(20700), shap=0.0, mean=19854.1, std=758.5];[name=Pressure(937.539), shap=0.5, mean=938.5, std=1.3];[name=PM1.0(2.06), shap=-0.1, mean=151.2, std=1103.6];[name=PM2.5(2.14), shap=0.0, mean=261.9, std=2214.4];[name=NC0.5(14.2), shap=-0.0, mean=777.6, std=5419.5];[name=NC1.0(2.215), shap=0.0, mean=287.6, std=2470.5];[name=NC2.5(0.05), shap=-0.0, mean=105.4, std=1157.8];

target=[name=smoke, prediction=no smoke, confidence=95%];

cf=[name=temperature(27.1)][name=humidity(47.2)][name=TVOC(1103.0)] [name=Raw Ethanol(19459.0)][name=Pressure(938.8)];

cf_pred=smoke

- **Output from Grammar**

The most important feature which influenced the prediction of no smoke with a confidence of 95% is the value of Pressure (938), which is lower than the mean value. The Pressure value is quite low which increases the chances of detecting smoke. The value of TVOC (44) is also an important element, because low values of TVOC means that the chances of detecting smoke is low, and the value of humidity (43), which is lower than the mean value also affects the prediction, because low values of humidity means that the chances of detecting smoke is low,. If Raw Ethanol was 19459, temperature was 27, NC2.5 was 0, humidity was 47 and Pressure was 938 then the classifier would have predicted smoke.

5.3.4. Occupancy

- **Grammar**

The following is the grammar used to create training outputs for stroke dataset.

S −> F SN T CF

F −>The main reason for the prediction O1 is the EF | The most important feature which influenced the prediction of O2 is the EF | The prediction of O2 has high dependence on EF | The most relevant factor for the prediction of O2 is the EF | The O2 outcome is primarily determined by the EF | The EF plays a significant role in O2 | Of primary importance for predicting O2 is the EF | The EF is important for predicting O2 | The EF is most important in predicting O2 | The EF is a dominant factor for the prediction of O2

SN −> . In addition , the EF also has a significant influence | . The EF is also an important element | . Moreover, the EF plays an important role | . The second important element is the EF | . Another important feature is the EF | . Furthermore, the EF has a considerable effect | . A second factor to consider is the EF | . The EF also has a significant effect | . The EF is another major factor

T −> and also the EF is relevant. | , while the third factor is the EF. | . Finally , the EF also influences the result .| and the EF also affects the prediction | and the EF also contributes to the result | The result is also affected by the EF | The EF is third factor that determines the outcome | It is important to mention the EF as well | The EF is also worth mentioning among the causes

O1 −> having vacancy | having ocupancy

O2 −> vacancy | occupancy

EF −> OTHER-NUM

OTHER-NUM −> value of f (v) DIST WHY

DIST −> , which is n-std standard deviations above the mean, | , which is n-std standard deviations below the mean, | , which is higher than the mean, | , which is lower than the mean,

CF −> FIRST-CF-F CF-F

FIRST-CF-F −> If CF-F-DESC

CF-F −> , CF-F-DESC |

CF-F-DESC −> OTHER-CF

OTHER-CF −> f was v

WHY −> The higher the value of EF, lower the chances of occupancy. | because when the values of EF improves the probability of occupancy also improves | The EF value is quite low which increases the chances of occupancy. | , because low values of EF reduces the probability of occupancy, | ϵ

- **Encoded Input**

```
input=[name=Temperature(19.39), @shap=-0.0,@ mean=20.6, std=1.0];
      [name=Humidity(31.0), shap=-0.0, mean=25.7, std=5.5];
      [name=Light(0.0), shap=0.1, mean=119.5, std=194.8];
      [name=CO2(434.0), shap=0.1, mean=606.5, std=314.3];
      [name=HumidityRatio(0.004314804), shap=0.0, mean=0.0, std
         =0.0];
target=[name=occupancy, prediction(vacancy), confidence=98%];
cf=[name=Temperature(19.5)][name=Humidity(27.0)][name=Light
   (289.8)][name=HumidityRatio(0.0)]; cf_pred=occupancy
```

- **Output from Grammar**

The most relevant factor for the prediction of vacancy with a confidence of 98% is the value of Light (0), which is lower than the mean value, because low values of Light reduces the probability of occupancy. Moreover, the value of CO2 (434), which is lower than the mean value, plays an important role, because low values of CO2 reduces the probability of occupancy, and also the value of HumidityRatio (0), which is 1 standard deviation above the mean is relevant. The higher the value of HumidityRatio, lower the chances of occupancy,. If Humidity was 27, HumidityRatio was 0, CO2 was 473 and Light was 289 the result would have been occupancy.

5.3.5. Heart Disease

- **Grammar**

The following is the grammar used to create training outputs for stroke dataset.

S −> F SN T CF

F −>The main reason for the prediction O1 is the EF | The most important feature which influenced the prediction of O2 is the EF | The prediction of O2 has high dependence on EF | The most relevant factor for the prediction of O2 is the EF | The O2 outcome is primarily determined by the EF | The EF plays a significant role in O2 | Of primary importance for predicting O2 is the EF | The EF is important for predicting O2 | The EF is most important in predicting O2 | The EF is a dominant factor for the prediction of O2

SN −> . In addition , the EF also has a significant influence | . The EF is also an important element | . Moreover, the EF plays an important role | . The second important element is the EF | . Another important feature is the EF | . Furthermore, the EF has a considerable effect | . A second factor to consider is the EF | . The EF also has a significant effect | . The EF is another major factor

T −> and also the EF is relevant. | , while the third factor is the EF. | . Finally , the EF also influences the result .| and the EF also affects the prediction | and the EF also contributes to the result | The result is also affected by the EF | The EF is third factor that determines the outcome | It is important to mention the EF as well | The EF is also worth mentioning among the causes

O1 −> having vacancy | having ocupancy

O2 −> vacancy | occupancy

EF −> AGE | OTHER-NUM | SEX | EIA | SLOP | THAL | TCP

AGE −> fact that patient is elderly WHY | fact that patient is young WHY | fact that the patient is middle-aged WHY

TCP −> typical angina type chest pain | atypical angina type chest pain | nonanginal pain type chest pain | asymptomatic type chest pain

SEX −> it is a male | it is a female

THAL −> thalium stress result that says normal | thalium stress result that says fixed defect | thalium stress result that says reversible defect

SLOP −> slope of the ST segment of peak exercise which in this case is flat | slope of the ST segment of peak exercise is upsloping | slope of the ST segment of peak exercise is downloping

EIA −> fact that the patient had a exercise induced angina | fact that the patient did not have exercise induced angina

OTHER-NUM −> value of f (v) DIST WHY

DIST −> , which is n-std standard deviations above the mean, | , which is n-std standard deviations below the mean, | , which is higher than the mean, | , which is lower than the mean,

CF −> FIRST-CF-F CF-F

FIRST-CF-F −> If CF-F-DESC

CF-F −> , CF-F-DESC |

CF-F-DESC −> OTHER-CF

SEX-CF −> it is a male | it is a female

THAL-CF −> thalium stress result showsnormal | thalium stress result shows fixed defect | thalium stress result shows reversible defect

SLOP-CF −> slope of the ST segment of peak exercise is flat | slope of the ST segment of peak exercise is upsloping | slope of the ST segment of peak exercise is downsloping

OTHER-CF −> f was v

WHY −> because low values of EF means that the chances of detecting heart disease with a confidence of PER% is low | The higher the value of EF, lower the chances of detecting heart disease. | because when the EF value increases, there are high chances that there is heart disease | The EF value is quite low which increases the chances of detecting heart disease.| ϵ

- **Encoded Input**

input=[name=age(43.0), shap=0.0, mean=54.4, std=9.1];[name=sex(Male), shap=-0.0];[name=type chest pain(asymptomatic), shap=-0.1];[name=maximum heart rate(181.0), shap=0.1, mean=149.7, std=23.2];[name=exercise induced angina(No), shap=0.0];[name=oldpeak(1.2), shap=-0.0, mean=1.0, std=1.1];[name=slope(flat), shap=-0.0];[name=number of colored vessels(0.0), shap=0.2, mean=0.7, std=0.9];[name=thalium(normal), shap=0.2];

target=[name=heart disease, prediction(heart disease), confidence=82%];

cf=[name=thalium(reversible defect)][name=age(48.0)][name=maximum heart rate(166.0)][name=oldpeak(0.5)];

cf_pred=no heart disease

- **Output from Grammar**

> The main reason for the prediction having a heart disease with a confidence of 82% is the thalium stress result that says normal. The second important element is the value of number of colored vessels (0), which is lower than the mean value, because low values of number of colored vessels means that the chances of detecting heart disease with a confidence of 82% is low, while the third factor is the value of maximum heart rate (181), which is 1 standard deviation above the mean, because when the maximum heart rate value increases, there are high chances that there is heart disease. If thalium stress result shows reversible defect, maximum heart rate was 166, oldpeak was 0 and age was 48 then the prediction would have been no heart disease.

5.3.6. Mammographic mass

- **Grammar**

The following is the grammar used to create training outputs for stroke dataset.

> S −> F SN T CF
>
> F −>The main reason for the prediction O1 is the EF | The most important feature which influenced the prediction of O2 is the EF | The prediction of O2 has high dependence on EF | The most relevant factor for the prediction of O2 is the EF | The O2 outcome is primarily determined by the EF | The EF plays a significant role in O2 | Of primary importance for predicting O2 is the EF | The EF is important for predicting O2 | The EF is most important in predicting O2 | The EF is a dominant factor for the prediction of O2
>
> SN −> . In addition , the EF also has a significant influence | . The EF is also an important element | . Moreover, the EF plays an important role | . The second important element is the EF | . Another important feature is the EF | . Furthermore, the EF has a considerable effect | . A second factor to consider is the EF | . The EF also has a significant effect | . The EF is another major factor

T −> and also the EF is relevant. | , while the third factor is the EF. | . Finally , the EF also influences the result .| and the EF also affects the prediction | and the EF also contributes to the result | The result is also affected by the EF | The EF is third factor that determines the outcome | It is important to mention the EF as well | The EF is also worth mentioning among the causes

O1 −> having no breast cancer | having breast cancer

O2 −> no breast cancer | breast cancer

EF −> OTHER-NUM | SHAPE | MARGIN | DENSITY

SHAPE −> irregular shape of the mass | round shape of the mass | oval shape of the mass | lobular shape of the mass

DENSITY −> high mass density |iso mass density | low mass density | fat-containing mass density

MARGIN −> mass margin which is microlobulated.| mass margin which is circumscribed | mass margin which is obscured | mass margin which is ill-defined | mass margin which is spiculated

OTHER-NUM −> value of f (v) DIST WHY

DIST −> , which is n-std standard deviations above the mean, | , which is n-std standard deviations below the mean, | , which is higher than the mean, | , which is lower than the mean,

CF −> FIRST-CF-F CF-F

FIRST-CF-F −> If CF-F-DESC

CF-F −> , CF-F-DESC |

CF-F-DESC −> OTHER-CF

SHAPE-CF −> shape of the mass was irregular | shape of the mass was round | shape of the mass was oval | shape of the mass was lobular

DENSITY −> mass density was high | mass density was iso | mass density was low | mass density fat-containing

MARGIN −> mass margin was microlobulated.| mass margin was circumscribed | mass margin was obscured | mass margin was ill-defined | mass margin was spiculated

OTHER-CF −> f was v

WHY −> The higher the value of EF, lower the chances of occupancy. | because when the values of EF improves the probability of occupancy also improves | The EF value is quite low which increases the chances of occupancy. | , because low values of EF reduces the probability of occupancy,

- **Encoded Input**

input=[name=BI-RADS(4), shap=-0.1, mean=4.4, std=1.9]; [name=Age(62), shap=0.1, mean=55.8, std=14.7]; [name=Shape(irregular), shap=0.1]; [name=Margin(microlobulated),shap=0.1]; [name=Density(low),shap=0.0];

target=[name=mass severity, prediction(malignant mass), confidence=80%];

cf=[name=Margin(ill-defined)][name=Age(63.0)];

cf_pred=benign mass

- **Output from Grammar**

The main reason for the prediction having breast cancer with a confidence of 80% is the irregular shape of the mass. The second important element is the fact that the patient is elderly, because when the Age value increases, there are high chances that mass is breast cancer, while the third factor is the the mass margin which is microlobulated. If the mass margin was ill-defined and Age was 63 then the prediction would have been no breast cancer.

5.3.7. Modified Cardio

- **Grammar**

The following is the grammar used to create training outputs for stroke dataset.

S −> F SN T CF

F −>The main reason for the prediction O1 is the EF |
The most important feature which influenced the prediction of O2 is the EF |
The prediction of O2 has high dependence on EF |
The most relevant factor for the prediction of O2 is the EF |
The O2 outcome is primarily determined by the EF |
The EF plays a significant role in O2 |
Of primary importance for predicting O2 is the EF |
The EF is important for predicting O2 |
The EF is most important in predicting O2 |
The EF is a dominant factor for the prediction of O2

SN −> . In addition , the EF also has a significant influence | . The EF is also an important element | . Moreover, the EF plays an important role | . The second important element is the EF | . Another important feature is the EF | . Furthermore, the EF has a considerable effect | . A second factor to consider is the EF | . The EF also has a significant effect | . The EF is another major factor

T −> and also the EF is relevant. | , while the third factor is the EF. | . Finally , the EF also influences the result .| and the EF also affects the prediction | and the EF also contributes to the result | The result is also affected by the EF | The EF is third factor that determines the outcome | It is important to mention the EF as well | The EF is also worth mentioning among the causes

P −> he | she

O1 −> having no cardiovascular disease | having a cardiovascular disease

O2 −> no cardiovascular disease | cardiovascular disease

EF −> AGE | OTHER-NUM | PA | ALCOHOL | GENDER| SMOKING | GLUCOSE | CHOLESTEROL

AGE −> fact that patient is elderly WHY | fact that patient is young WHY | fact that the patient is middle-aged WHY

PA −> physical activity of the patient WHY | inactivity of the patient WHY

ALCOHOL −> use of alcohol WHY | absence of use of alcohol WHY

SMOKING −> fact that the patient is a smoker WHY | fact that the patient is not a smoker WHY

GLUCOSE −> v level of glucose WHY

CHOLESTEROL −> v level of cholesterol WHY

OTHER-NUM −> value of f (v) DIST WHY

DIST −> , which is n-std standard deviations above the mean, | , which is n-std standard deviations below the mean, | , which is higher than the mean, | , which is lower than the mean,

CF −> FIRST-CF-F CF-F

FIRST-CF-F −> If CF-F-DESC

CF-F −> , CF-F-DESC |

CF-F-DESC −> ALCOHOL-CF | SMOKE-CF | GENDER-CF | OTHER-CF

ALCOHOL-CF −> the patient drank alcohol | the patient did not drink alcohol

SMOKE-CF −> the patient was a smoker | the patient was not a smoker

GENDER-CF −> the patient was a male | the patient was a female

OTHER-CF −> f was v

WHY −> , where high values of this attribute are associated with a high probability of cardiovascular disease | , where high values of this attribute are associated with a low probability of cardiovascular disease | , where low values of this attribute are associated with a high probability of cardiovascular disease | , where low values of this attribute are associated with a low probability of cardiovascular disease , |

- **Encoded Input**

input= [name=age(52.271232876712325), shap=0.0, mean=53.3, std=6.8];[name=gender(Male), shap=0.0];[name=systolic blood pressure(14), shap=0.2, mean=126.6, std=16.6];[name=diastolic blood pressure(90), shap=-0.0, mean=81.3, std=9.3];[name=cholesterol(Normal), shap=0.0];[name=glucose(Normal), shap=-0.0];[name=smoking(No), shap=-0.0];[name=alcohol(No), shap=-0.0];[name=physical activity(Yes), shap=0.0];[name=BMI(21.92612582222972), shap=0.0, mean=27.4, std=5.0];

target= [name=cardiovascular disease, prediction=no disease, confidence=98%];

cf= [name=age(61.9)][name=diastolic blood pressure(80.0)][name=BMI(26.1)];

cf_pred=cardiovascular disease

- **Output from Grammar**

> The first element which influenced the prediction of no disease with a confidence of 98% is, the value of systolic blood pressure (14), which is lower than the mean. Moreover, the normal level of cholesterol plays an important role. Finally, the fact that the patient is middle-aged also influences the result, where low values of this feature are associated with a low probability of cardiovascular disease,. If diastolic blood pressure was 80, BMI was 26 and age was 61 the result would have been cardiovascular disease.

5.4. Modified encoding

We introduced a new encoding with information about the probability of the prediction to give the users more insights. As the previous encoding was long and thus was not able to train for datasets with many features we tried to change the encoding to shorten the length. The form of giving the input values was changed from name=age, value=52.0 to name=age(52.0)

- **Old Encoding**

input=[name=age, value=52.271232876712325, shap=0.0, mean=53.3, std=6.8];[name=gender, value=Male, shap=0.0];[name=systolic blood pressure, value=14, shap=0.2, mean=126.6, std=16.6];[name=diastolic blood pressure, value=90, shap=-0.0, mean=81.3, std=9.3];[name=cholesterol, value=Normal, shap=0.0];[name=glucose, value=Normal, shap=-0.0];[name=smoking, value=No, shap=-0.0];[name=alcohol, value=No, shap=-0.0];[name=physical activity, value=Yes, shap=0.0];[name=BMI, value=21.92612582222972, shap=0.0, mean=27.4, std=5.0];
prediction=no disease;

cf=[name=age, value=61.9][name=diastolic blood pressure, value=80.0][name=BMI, value=26.1];
cf_pred=cardiovascular disease

- **New Encoding**

input= [name=age(52.271232876712325), shap=0.0, mean=53.3, std=6.8];[name=gender(Male), shap=0.0];[name=systolic blood pressure(14), shap=0.2, mean=126.6, std=16.6];[name=diastolic blood pressure(90), shap=-0.0, mean=81.3, std=9.3];[name=cholesterol(Normal), shap=0.0];[name=glucose(Normal), shap=-0.0];[name=smoking(No), shap=-0.0];[name=alcohol(No), shap=-0.0];[name=physical activity(Yes), shap=0.0];[name=BMI(21.92612582222972), shap=0.0, mean=27.4, std=5.0];

target= [name=cardiovascular disease, prediction=no disease, confidence=98%];

cf= [name=age(61.9)][name=diastolic blood pressure(80.0)] [name=BMI(26.1)];

cf_pred=cardiovascular disease

5.5. Preventing overfitting on one domain

There were more instances of cardio and stroke datasets which lead to model overfitting on them. To handle the overfitting of the models on cardio or stroke datasets, we experimented by taking a subset of 50 instances from each dataset to see the effect on the model results.

The final model outputs from both are reported below.

The subset was taken with the following composition:

- 50 instances from cardio
- 50 instances from stroke
- 50 instances from occupancy
- 50 instances from breast cancer

The result of this experiment showed that the model trained with the subset performed better at generalizing for a new unseen dataset than the model trained with the whole dataset.

You can see the outputs of both models on the diabetes test case below.

Trained with whole dataset

In the output we got from the model trained with the whole dataset, you can see that the model says diastolic blood pressure instead of diabetes. The model is clearly overfitting on the cardio dataset as the explanation is for the cardiovascular domain rather than diabetes.

> The first element which influenced the prediction of diastolic blood pressure (35), which is 1 standard deviation above the mean, is the value of glucose (148), which is 1 standard deviation above the mean. The second important element is the value of BMI (33), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of cardiovascular disease, and the fact that the patient is elderly also affects the prediction, where high values of this feature are associated with a high probability of cardiovascular disease. If glucose was 165, BMI was 31 and age was 49 then the classifier would have predicted no diabetes.

Trained with 50 instances each

The result from the model trained on the subset is given below. You can see a clear difference in the outputs as this output correctly gives an explanation for the diabetes domain. This model better generalizes the result than the model trained on the whole dataset.

> The first element which influenced the prediction of diabetes with a confidence of 95% is the fact that the patient is young, where low values of this feature are associated with a high probability of diabetes. Moreover, the fact that the patient is middle-aged plays an important role, where low values of this feature are associated with a high probability of diabetes. If age was 49, BMI was 26 and blood pressure was 68 then the prediction would have been no diabetes.

5.6. Improving the Diversity of the explanations

As we trained the model with the explanations generated with the grammar, the model outputs are very much similar. We explored different methods to make the generated output more diverse using the capabilities of T5. Using the outputs from the model trained on the subset we tried different approaches to generate more diverse explanations for the prediction.

Paraphrasing

Modifying a text without changing its meaning is called paraphrasing. We fine-tuned T5 on the TaPaCo dataset [18] for paraphrasing. Tapaco is a paraphrase corpus that can be used to train models for paraphrasing. We use only paraphrases in the English language. We used the prefix "paraphrase:" for model training. The sample input and the output from the paraphrase model are given below.

Sample Input

Model output from the stroke domain was taken to give as input to this model.

> The most relevant factor for the prediction of low probability of getting stroke with a confidence of 65% is, the value of BMI (43.8), which is 2 standard deviations above the mean, where high values of this feature are associated with a low probability of getting stroke The second important element is the fact that the patient was self-employed. Finally, the fact that the patient did not have hypertension also influences the result. If the patient worked in a private company, the patient lived in a urban area, BMI was 37 and average glucose level was 221 then the classifier would have predicted high probability of getting stroke

Sample Output

As you can see in the output, the model uses the correct feature names and values from the input. It changes the phrases with new meaningful phrases. Also, note that the output is shorter as it has emitted information about the second and the third important features. Also, it replaced the word "if" with "since" which in this case, does not give the intended meaning to the sentence.

> The major factor for a 65% probability of getting stroke is to the value of BMI (43.8), who is 2 standard deviations from the mean, where high values of this feature are

associated with a low likelihood of getting stroke. Since the patient lived in a suburb, BMI was 37 and glucose level was 221, then the classifier would have predicted high probability of getting stroke.

Sample Input

We give another input from the cardio domain.

The most relevant factor for the prediction of presence of cardiovascular disease with a confidence of 84% is, the value of systolic blood pressure (130). The second important element is the value of diastolic blood pressure (90), where low values of this feature are associated with a high probability of cardiovascular disease, while the third factor is the value of BMI (27.8). If diastolic blood pressure was 80, BMI was 27 and age was 53 the result would have been no disease."

Sample Output

In this output, you can see that the model hallucinates new words like 68 years, which is not present in the input.

The most relevant factor for the prediction of early detection of cardiovascular disease with a confidence of 68 years is the value of systolic blood pressure, 130. The second key element is the value of diastolic blood pressure 90, where the low values of this feature's associated with a high probability of cardiovascular disease, while the third factor is the value of BMI (27.8). If diastolic blood pressure was 80, age was 53 and it would have been no disease

This model was not trained with any sample paraphrases relating to our explanations. Fine-tuning, the model with more example paraphrases close to our domains can make the performance much better.

Summarizing

The T5 model is pre-trained on multiple tasks. One of the tasks is summarization. We investigated this usecase of T5 on our system by giving the generated output as an input to T5 to summarize by using the prefix "summarize:"

Sample Input

We gave output from the stroke domain as input for summarization

The most relevant factor for the prediction of low probability of getting stroke with a confidence of 65% is, the value of BMI (43.8), which is 2 standard deviations above the mean, where high values of this feature are associated with a low probability of getting stroke The second important element is the fact that the patient was self-employed. Finally, the fact that the patient did not have hypertension also influences the result. If the patient worked in a private company, the patient lived in a urban area, BMI was 37 and average glucose level was 221 then the classifier would have predicted high probability of getting stroke

Sample Output

In the given output, you can see that the model summarized the input correctly. But note that the output omits the feature values and some counterfactuals were omitted too.

a high BMI is associated with a low probability of getting stroke. the fact that the patient was self-employed also influences the result. a high glucose level also influences the prediction of low stroke probability. a patient who had no hypertension is not likely to get stroke.

Sample Input

We gave another sample input from the cardio domain to summarize.

The most relevant factor for the prediction of presence of cardiovascular disease with a confidence of 84% is, the value of systolic blood pressure (130). The second important element is the value of diastolic blood pressure (90), where low values of this feature are associated with a high probability of cardiovascular disease, while the third factor is the value of BMI (27.8). If diastolic blood pressure was 80, BMI was 27 and age was 53 the result would have been no disease."

Sample Output

Also in this example, you can see that the model output reduces the amount of information in the explanation. Here, it omitted the third important factor BMI from the output.

systolic blood pressure is the most important factor for cardiovascular disease prediction. value of diastolic blood pressure is also important for predicting risk of cardiovascular disease. low values of this feature are associated with a high probability of cardiovascular disease. if systolic blood pressure was 80, BMI was 27 and age was 53 the result would have been no disease.

Summarization brings diversity to the outputs. But it also omits the feature values from its outputs which makes the explanation less detailed.

Repetition Penality

The repetition penalty(RP) parameter is for avoiding repetitions in the generated texts. The default value is 1.0. The value can range from 1.0 to infinity.

We experimented by changing the repetition penalty value while generating the outputs. Some sample outputs with different RP values are given below. The example is shown for a sample input from the diabetes dataset, which was not used during training.

RP = 1.0 (default)

When the value is 1.0, the model generates the output with no penalty.

> The first element which influenced the prediction of diabetes with a confidence of 85% is, the value of glucose (148), which is 1 standard deviation above the mean. The second important element is the value of BMI (33.5), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of high probability of getting diabetes. Finally, the value of glucose (125), which is 1 standard deviation above the mean also influences the result, where high values of this feature are associated with a high probability of high probability of getting diabetes. If age was 49 and BMI was 33 then the prediction would have been no diabetes.

RP = 1.5

When the value is increased to 1.5, the output seems less meaningful but it uses phrases not given in our grammar. It introduces new words - insulin instead of glucose. Also, note that the model uses cholesterol which is not related to our input.

> The first element which influenced the prediction of diabetes with a confidence of 85% is, the value of glucose (148), which is 1 standard deviation above the mean. The second important element is the value of BMI (33.5), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of high probability of getting diabetes and also the value of cholesterol (164) is relevant. If age was 49 and insulin was 168 then the prediction would have been no diabetes. "It was a high probability of getting diabetes

RP = 20

When the value is increased the output becomes more and more diverse. But the level of hallucination also increases.

> The first element which influenced the prediction of diabetes with 85% is, how much blood pressure (72), where low values are associated. The second important element for this feature set includes: value and also inactivity on both sides; when high values occur then the classifier will have predicted no diabetes or age-49 as well! If BMI was 33 and glucose level was 64 then would have been no diabetes? "No diabetes"!" said one person who has been diagnosed by age at 49 but still had to wait until age was 50) it would have been No Diabetes

We can see from the examples that changing the RP values makes the input more diverse. It can make the model generate outputs that are more conversation-like. But this in turn also increases the hallucinations affecting the meaning of the explanations.

6 | Experiments

6.1. Testing the old model on a new domain

As the first approach to the thesis, we experimented with testing the two models - first - trained on cardio, and second - trained on cardio, and fine-tuned on diabetes, with a new dataset of breast cancer.

The result showed that the model fails to generate meaningful output without fine-tuning.

As you can see in figure 6.1, the output is overfitting on the cardio domain. The model was only trained on cardio, so it fails to generalize on any other domain.

Explanation 1

The first element which influenced the prediction of Benign is the value of area_worst (314.9000), which is lower than the mean, where low values of this feature are associated with a low probability of Malignant. In addition, the value of concave points_worst (0.0623), which is lower than the mean, also has a significant influence, where low values of this feature are associated with a low probability of Malignant, and also the value of perimeter_worst (65.1300), which is lower than the mean is relevant, where low values of this feature are associated with a low probability of Malignant,. If radius_mean was 10.26, texture_mean was 12.22, perimeter_mean was 65.75, area_mean was 321.6, smoothness_mean was 0.09996, compactness_mean was 0.07542, concavity_mean was 0.01923, concave points_mean was 0.01968, symmetry_mean was 0.18, fractal_dimension_mean was 0.06569, radius_se was 0.1911, texture_se was 0.5477, perimeter_se was 1.348, area_se was 11.88, smoothness_se was 0.005682, compactness_se was 0.01365, concavity_se was 0.008496, concave points_se was 0.006929, symmetry_se was 0.01938, fractal_dimension_se was 0.002371, radius_worst was 11.38, texture_worst was 15.65, perimeter_worst was 73.23, area_worst was 394.5, smoothness_worst was 0.1343, compactness_worst was 0.165, concavity_worst was 0.08615, concave points_worst was 0.06696, symmetry_worst was 0.2937 and fractal_dimension_worst was 0.07722 the opposite would have been predicted.

Explanation 2

The first element which influenced the prediction of presence of disease is the value of systolic blood pressure (150), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of cardiovascular disease. The second important element is the well above normal level of cholesterol. The fact that the patient is middle-aged is also worth mentioning among the causes, where high values of this feature are associated with a high probability of cardiovascular disease,. If systolic blood pressure was 140, age was 57 and BMI was 27 then the prediction would have been no disease.

Figure 6.1: Output from old GPT-2 model trained on cardio

model fine-tuned on diabetes

The output from the model trained on cardio and fine-tuned on diabetes on breast cancer input is given below. As you can see, even this model fails to generate meaningful output for any new domain.

Explanation 1

The most relevant factor for the prediction of Benign is the value of smoothness_worst (0.0895), which is lower than the mean. The value of area_mean (629.9000) is also an important element and also the value of area_worst (709.0000) is relevant. If radius_mean was 10.57, texture_mean was 20.22, perimeter_mean was 70.15, area_mean was 338.3, smoothness_mean was 0.09073, compactness_mean was 0.166, concavity_mean was 0.228, concave points_mean was 0.05941, symmetry_mean was 0.2188, fractal_dimension_mean was 0.0845, radius_se was 0.1115, texture_se was 1.231, perimeter_se was 2.363, area_se was 7.228, smoothness_se was 0.008499, compactness_se was 0.07643, concavity_se was 0.1535, concave points_se was 0.02919, symmetry_se was 0.01617, fractal_dimension_se was 0.0122, radius_worst was 10.85, texture_worst was 22.82, perimeter_worst was 76.51, area_worst was 351.9, smoothness_worst was 0.1143, compactness_worst was 0.3619, concavity_worst was 0.603, concave points_worst was 0.1465, symmetry_worst was 0.2597 and fractal_dimension_worst was 0.12 then the prediction would have been Malignant.

Explanation 2

The first motivation for the prediction of presence of disease is the value of BMI (29), which is lower than the mean, where low values of this feature are associated with a low probability of diabetes. The fact that the value of pregnancies is 1 is also an important element. Finally, the value of age (28) also influences the result. If pregnancies was 1. 0, BMI was 29. 0, diabetes pedigree function was 0. 0, skin thickness was 0. 0, blood pressure was 75. 0 and glucose was 92. 0

Figure 6.2: Output from old model finetuned on diabetes

So we moved to retrain the model with a third dataset to see how it performs on a third unseen dataset

6.2. Fine-tuning GPT2

- adding new tokens

  When fine-tuning GPT-2, we experimented with adding feature names as special tokens. We introduced the features with more than one word like "systolic blood pressure" in two ways. One with spaces like "systolic blood pressure" and another with three different words like "systolic", "blood", and "pressure". You can see from figure 6.3, that the one trained without spaces performs better than the one with spaces.

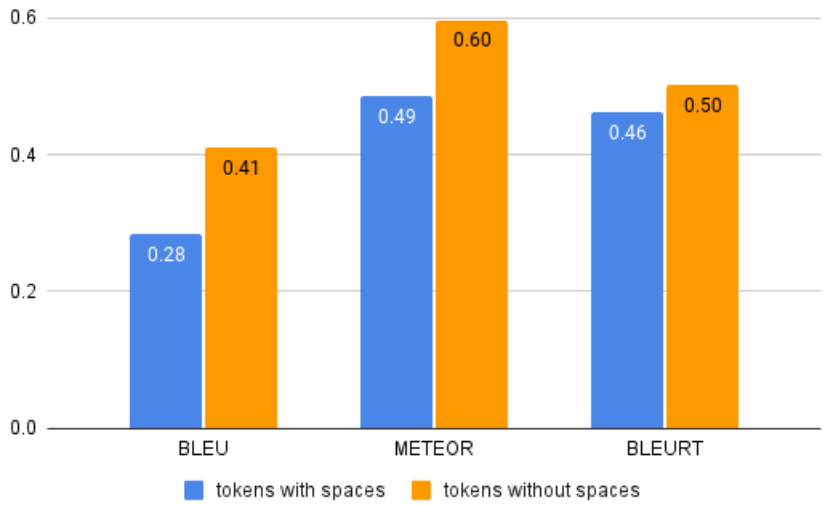

Figure 6.3: models trained with feature names as tokens, with and without spaces between words

- Cardio + Stroke epochs

  The GPT-2 model was trained with the cardio and stroke datasets. 7000 Values from cardio were taken for the training.

  The first graph shows the performance of the model when trained for different epochs.

  We can see that the evaluation values do not improve after two epochs. So we select this model for testing.

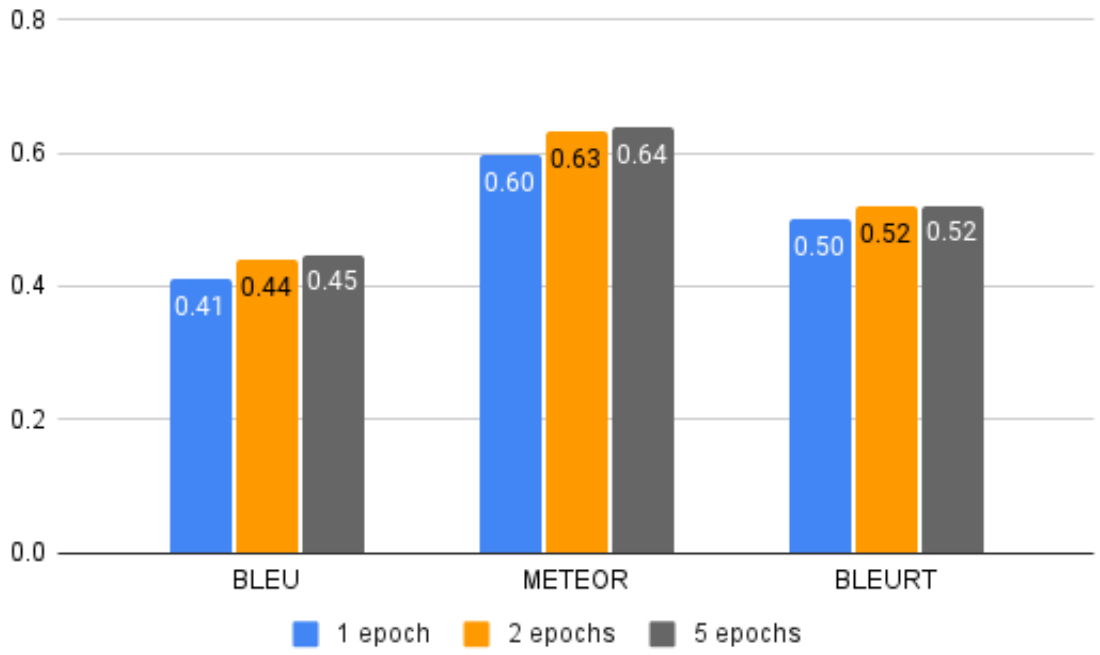

Figure 6.4: cardio + stroke model tested with different epochs

- Cardio + Stroke tests

  We test the model trained on cardio and stroke, on cardio, stroke, and diabetes datasets to analyze the generalization capabilities. The result on diabetes will tell us about the generalization performance. The model performs well for stroke and cardio inputs but the performance on the diabetes dataset is not great. But this is expected, as the diabetes input is a new domain for the model.

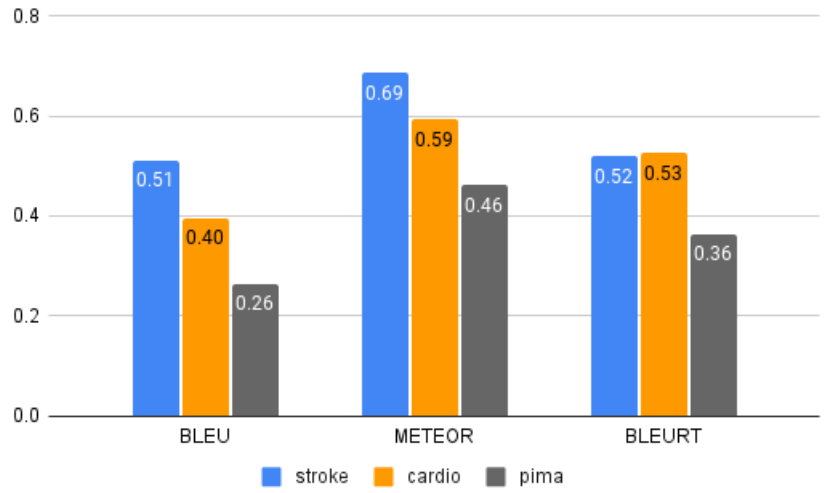

Figure 6.5: model trained on cardio and stroke

- Cardio + pima epochs Then the GPT-2 model was trained with cardio and Pima dataset. 7000 Values from cardio were taken for the training.

  The graph shows the performance of the model when trained for different epochs.

  We can see that the evaluation values do not improve after two epochs even here. So we select this model for testing.

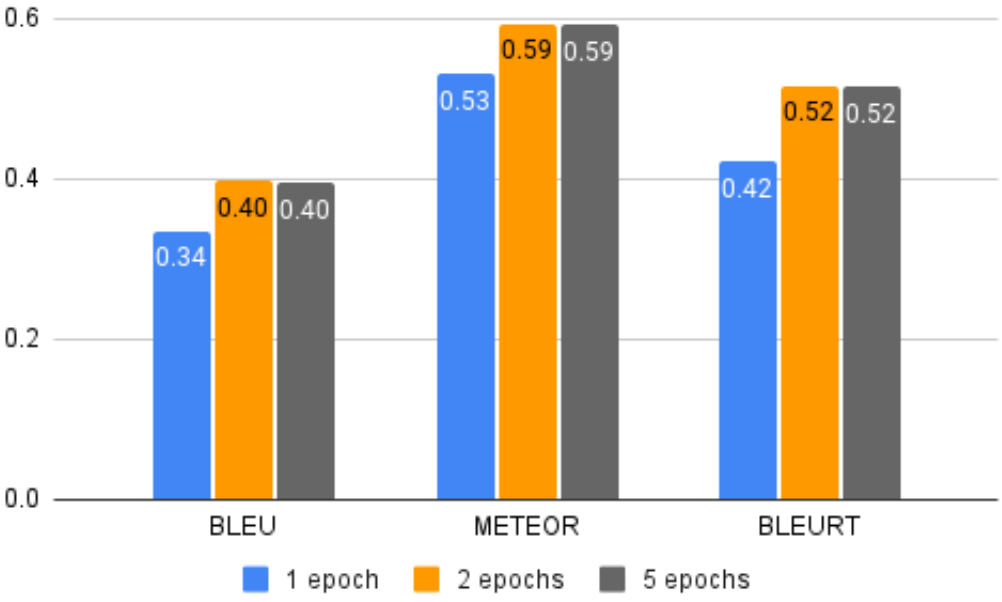

Figure 6.6: cardio + pima model tested with different epochs

- Cardio + Pima tests

  We test the model trained on cardio and Pima, on cardio, stroke, and diabetes datasets to analyze the generalization capabilities. The result of the stroke dataset will tell us about the generalization performance. The model performs well for Pima and cardio inputs but the performance on the stroke dataset is low.

  Comparing figures 6.5 and 6.7, The performance of this model on stroke is lower than the performance of cardio + stroke on diabetes.

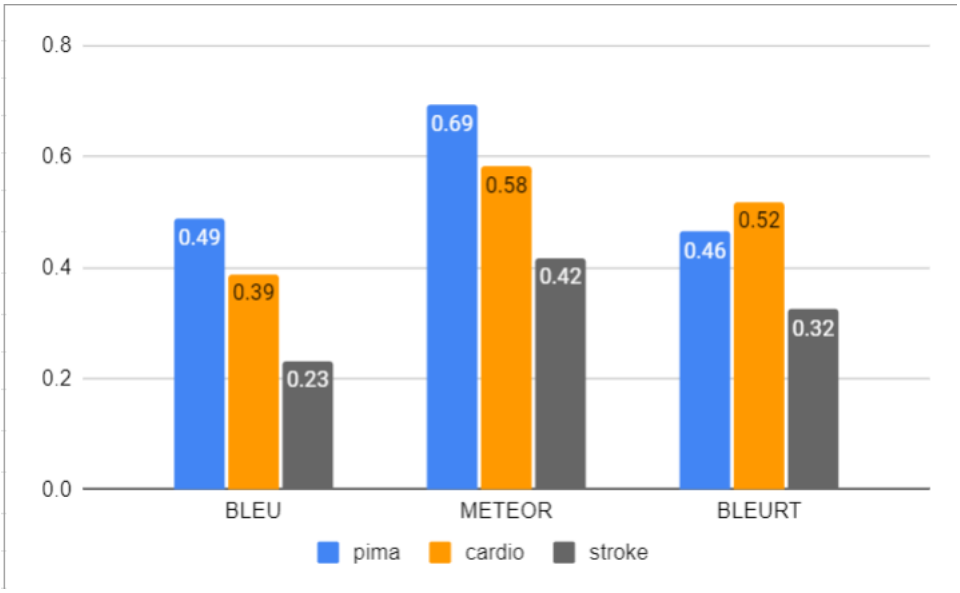

Figure 6.7: testing the model trained on cardio and pima-diabetes

- stroke + Pima epochs

  Then the GPT-2 model was trained, with stroke and Pima dataset.

  The graph shows the performance of the model when trained for different epochs.

  Here the model performance improves considering BLEU and BLEURT scores in the 5th epoch. So we select the model trained for 5 epochs model for testing.

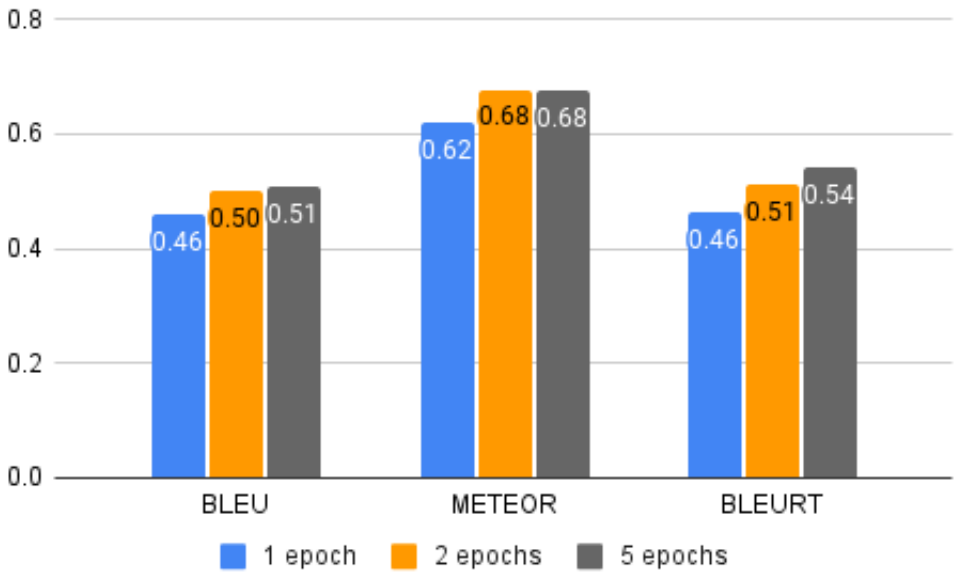

Figure 6.8: model trained on Pima and stroke

- stroke + Pima tests

  We test the model trained on stroke and pima, on cardio, stroke, and diabetes datasets to analyze the generalization capabilities. The result on cardio will tell us about the generalization performance. The model performs well for pima and stroke inputs but the performance on the cardio dataset is low.

  Comparing figures 6.5, 6.7, and 6.8, The performance of this model is best in generalizing. The model is able to perform better on an unseen domain.

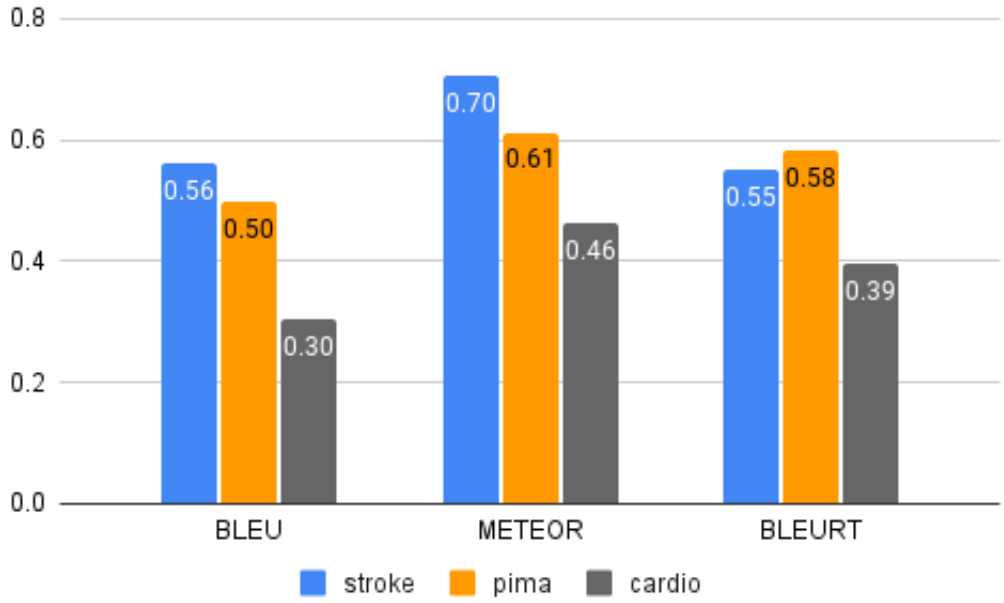

Figure 6.9: Pima + stroke model test with different epochs

6.3. T5 vs GPT-2

In this part, we show the result of the comparison between the T5 and GPT2 models.

Figure 6.10 shows the result of the experiment done by training both T5 and GPT-2 models with the cardio+ stroke dataset. GPT-2 model was trained for 5 epochs and T5 for 2 epochs. Even with less training, the T5 model performs better showing better BLEU and METEOR scores.

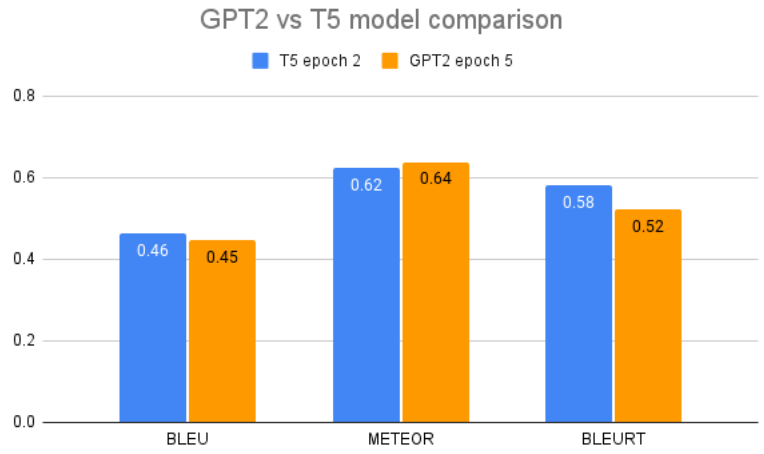

Figure 6.10: GPT-2 vs T5 model comparison

Here we compare the result of GPT-2 trained for 5 epochs on T5 trained for 5 Epochs. Here T5 performs better in all metrics indicating T is better than GPT-2 for our task-

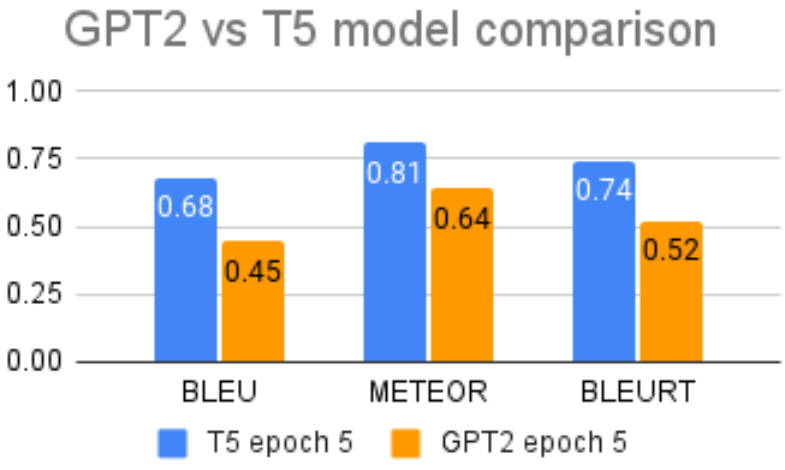

Figure 6.11: GPT-2 vs T5 model comparison

As the T5 model shows better performance, we decided to use T5 for further experiments. This also shows that the textual explanation task is better modeled as a translation task as T5 is an encoder-decoder model pre-trained for translation.

6.4. Multi-domain training and testing

Here we run the same training and testing done in section 6.2, by fine-tuning T5 instead of GPT-2

- Cardio + Stroke dataset

  First, we train T5 with cardio + stroke for different epochs. The best result is given by epoch 5, so we use this for further experiments.

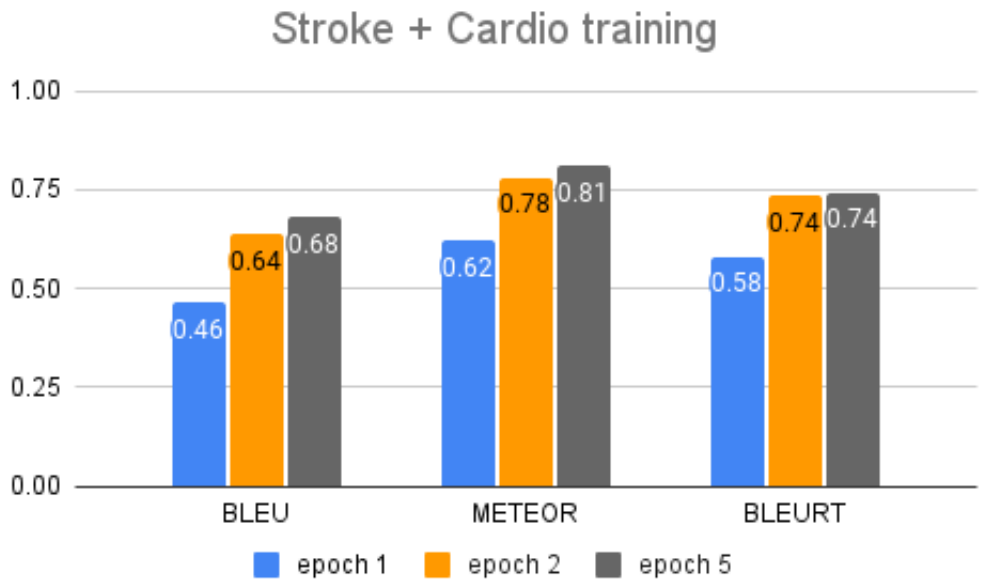

Figure 6.12: Stroke + cardio model

Here we test the model on the diabetes dataset to see the generalizing capability.

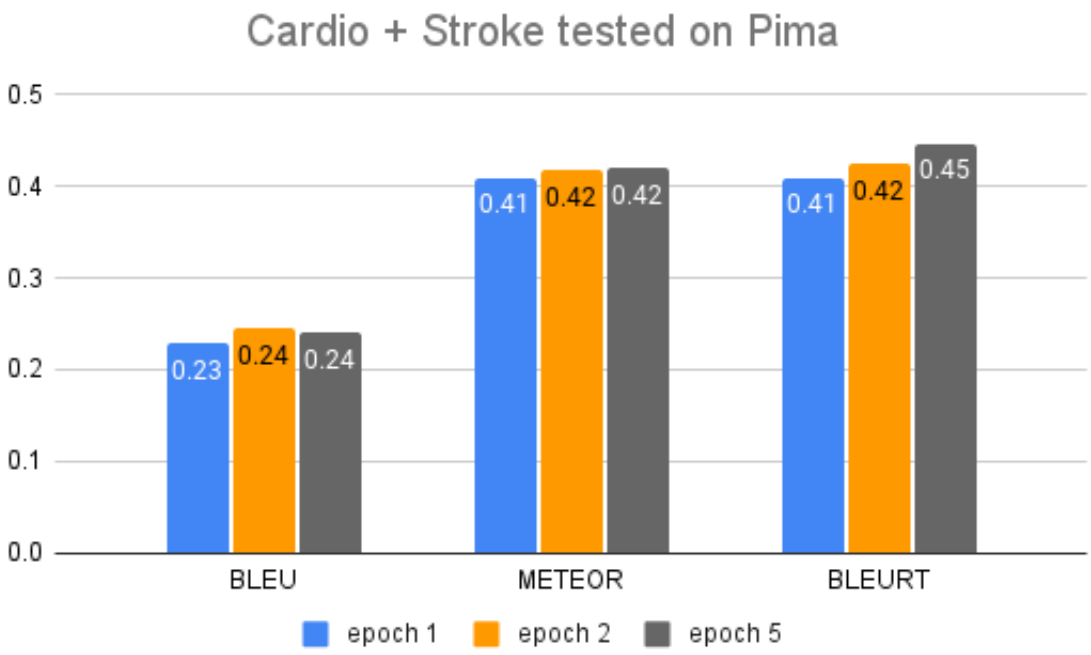

Figure 6.13: Cardio + stroke model tested on diabetes

The best result was obtained from a model trained for 5 epochs on cardio and stroke. On a new domain, it gives a good BLEURT score.

- occupancy dataset

  Here we train the model on a new dataset. We use the occupancy dataset for this test. We train it for two epochs. The figure in 6.14 shows that the performance reduces in epoch 2. So we train it on only one epoch. This gives an insight that when we use only one domain, the model tends to perform better with just one epoch. We get to this conclusion as the models used in [19] were trained for only one epoch both for cardio and when fine-tuned on diabetes

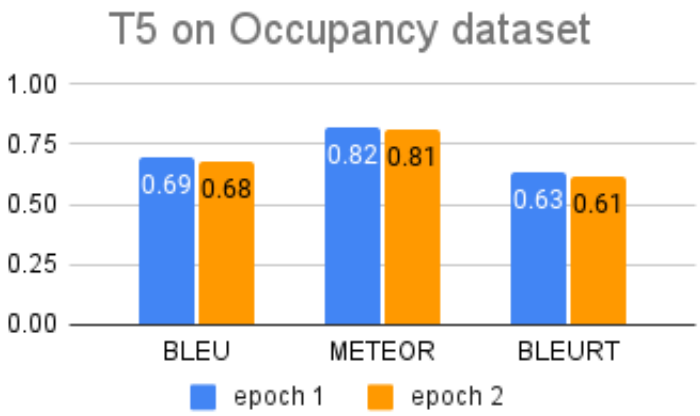

Figure 6.14: model trained only on occupancy dataset

- Multi-domain training with CSOB dataset Here we train the T5 model with cardio + stroke + occupancy + breast cancer dataset (CSOB). We train the model for 10 epochs.

  We got the best result in epoch 8. So we use this model for further experiments.

  We can see that this model provides the best BLUE and METEOR scores than any model tested so far.

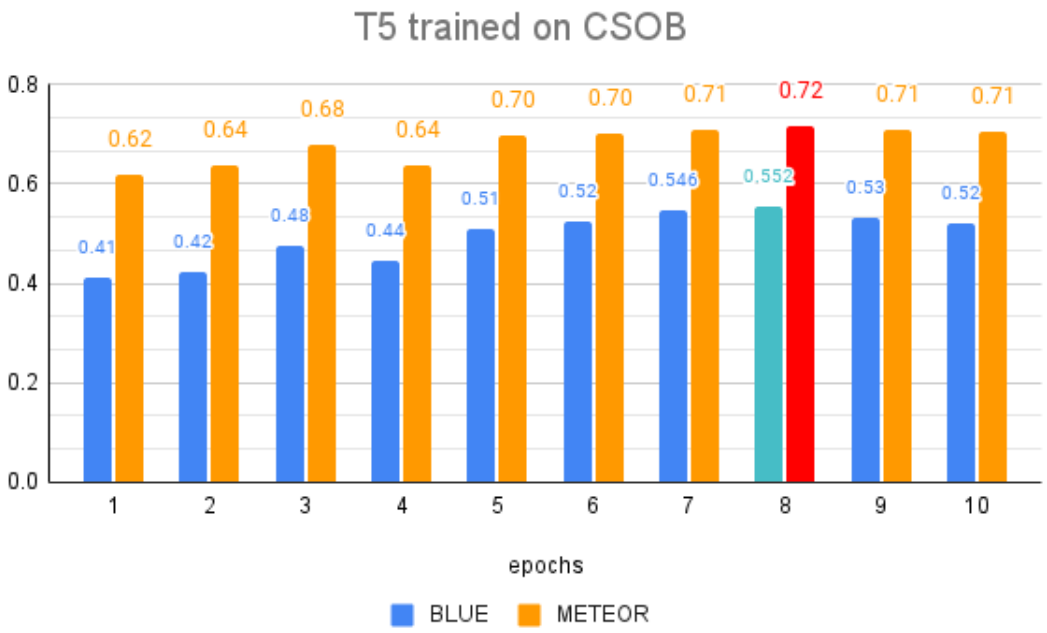

Figure 6.15: CSOB-model trained with cardio, stroke, occupancy, and breast cancer datasets

Here let us test the generalizing capability of these models by testing it on the diabetes dataset. The model trained for six epochs gives the best score on the new domain.

Note that the BLUE and METEOR score of the model trained for 6 epochs is the best result obtained for far on the unseen domain, here the diabetes dataset. So this is the best generalizing model so far in the experiment.

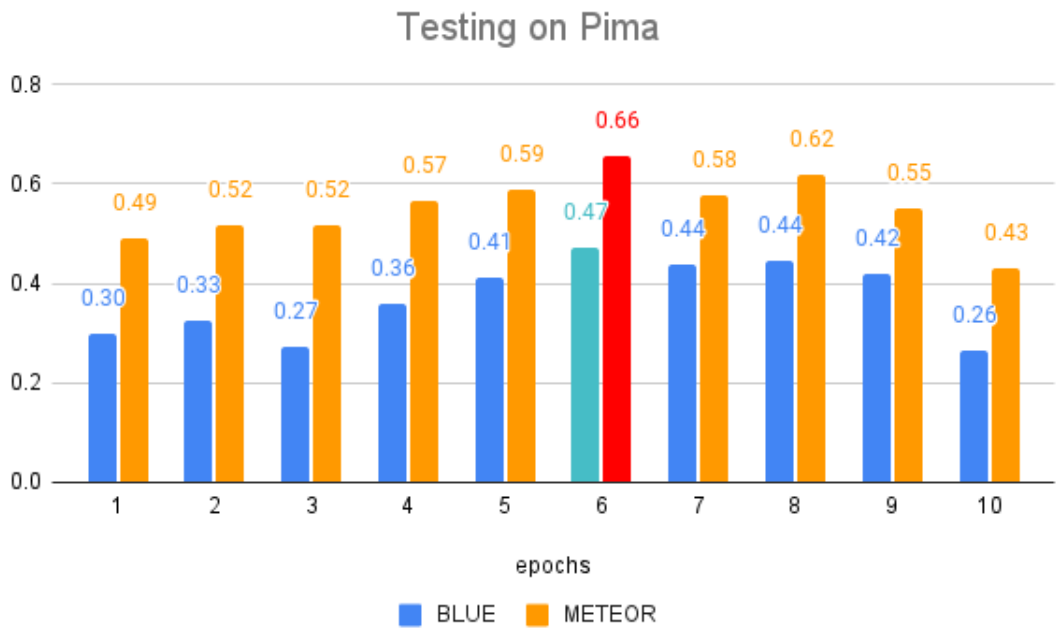

Figure 6.16: CSOB-model tested on pima

With these experiments, we get to the conclusion that, when introducing new datasets from new different domains, the performance of the model in generalizing for new domains improves.

6.5. Encodings

The inputs are given to the model in an encoded form. We need this encoding to include information important for explanations. These are the two encodings we tried.

Encoding 0

This was the initial encoding we tried. With this encoding, the model was not correctly using the prediction name while giving explanations. So we needed a new encoding.

input=[name=age(84), shap=0.1, mean=33.3, std=11.8];[name=pregnancies(6), shap=0.0, mean=3.9, std=3.4];[name=glucose(148), shap=0.2, mean=120.6, std=29.9];[name=blood pressure(72), shap=0.0, mean=72.4, std=11.5];[name=skin thickness(35), shap=0.0, mean=28.8, std=8.6];[name=insulin(125), shap=0.0, mean=130.5, std=55.4];[name=BMI(33), shap=0.0];[name=diabetes pedigree function(0.627), shap=0.0, mean=0.5, std=0.3];

target=diabetes;

cf=[name=age(49.0][name=glucose(165.0][name=blood pressure(68.0)][name=skin thickness(26.0)][name=insulin(168.0)];

cf_pred=no diabetes

New encoding We change the encoding by including the confidence the model has on the prediction. To add this, we change the way the target is encoded.

input=[name=age(84), shap=0.1, mean=33.3, std=11.8];[name=pregnancies(6), shap=0.0, mean=3.9, std=3.4];[name=glucose(148), shap=0.2, mean=120.6, std=29.9];[name=blood pressure(72), shap=0.0, mean=72.4, std=11.5];[name=skin thickness(35), shap=0.0, mean=28.8, std=8.6];[name=insulin(125), shap=0.0, mean=130.5, std=55.4];[name=BMI(33), shap=0.0];[name=diabetes pedigree function(0.627), shap=0.0, mean=0.5, std=0.3];

target=[name=diabetes, prediction(diabetes), confidence=55%];

cf=[name=age(49.0)][name=glucose(165.0)][name=blood pressure(68.0)][name=skin thickness(26.0)][name=insulin(168.0)];

cf_pred=no diabetes

6.6. T5 model explanations

In this section, we show the results we got from fine-tuned T5 model on different datasets.

6.6.1. Results of CSOB

In the following section, we give the outputs we got from the model trained for different epochs on CSOB datasets. Note that CSOB refers to the cardio + stroke + occupancy + breast cancer dataset. We give results we got from testing it on cardio, stroke, occupancy, and breast cancer, which shows how well the model performs in the training domains. Then we give the output we got when we tested it on the diabetes dataset. It will show us how well the models can generalize on an unseen domain.

Cardio

By examining the results on cardio, we can see that the model with epochs 8 and 10 gave the same results showing that no further training is required. Also after epoch 5, the third factor changed from diastolic blood pressure to cholesterol, indicating better training results.

ep	Outputs
1	The first element which influenced the prediction of no disease with a confidence of 78% is, the value of systolic blood pressure (120). Moreover, the fact that the patient is middle-aged plays an important role, where low values of this feature are associated with a low probability of cardiovascular disease, while the third factor is the value of diastolic blood pressure (80). If the patient was a female, age was 43 and BMI was 24 then the classifier would have predicted cardiovascular disease. If age was 41 then the classifier would have predicted cardiovascular disease.
5	The first element which influenced the prediction of no disease with a confidence of 78% is, the value of systolic blood pressure (120). Moreover, the fact that the patient is elderly plays an important role, where low values of this feature are associated with a low probability of cardiovascular disease, while the third factor is the value of diastolic blood pressure (80). If age was 41, BMI was 24 and age was 41 then the prediction would have been cardiovascular disease
8	The first element which influenced the prediction of no disease with a confidence of 78% is, the value of systolic blood pressure (120). Moreover, the fact that the patient is middle-aged plays an important role, where low values of this feature are associated with a low probability of cardiovascular disease, while the third factor is the normal level of cholesterol. If age was 41 and BMI was 24 then the prediction would have been cardiovascular disease.
10	The first element which influenced the prediction of no disease with a confidence of 78% is, the value of systolic blood pressure (120). Moreover, the fact that the patient is middle-aged plays an important role, where low values of this feature are associated with a low probability of cardiovascular disease, while the third factor is the normal level of cholesterol. If age was 41, BMI was 24 and age was 41 then the prediction would have been cardiovascular disease.

Stroke

In the stroke domain test result, we can see that epoch 8 and 10 gives the same results. The model with 8 epochs uses different phrases than the one with 5 epochs. The model with 5 epochs takes more correct features than the one with 1 epoch.

ep	Outputs
1	The first element which influenced the prediction of low probability of getting stroke with a confidence of 86% is, the fact that the patient is self-employed, where low values of this feature are associated with a low probability of getting stroke The second important element is, the fact that the patient did not have any heart disease, where low values of this feature are associated with a low probability of getting stroke If the patient was a smoker, average glucose level was 22 and age was 57 then the classifier would have predicted high probability of getting stroke.
5	The first element which influenced the prediction of low probability of getting stroke with a confidence of 86% is, the fact that the patient was self-employed. Moreover, the fact that the patient was married plays an important role and also the fact that the patient was self-employed is relevant. If the patient lived in a rural area, average glucose level was 221, age was 57 and BMI was 37 then the prediction would have been high probability of getting stroke
8	The first element which influenced the prediction of low probability of getting stroke with a confidence of 86% is, the fact that the patient was self-employed. Moreover, the fact that the patient was married plays an important role and the fact that the patient was married also affects the prediction. If the patient lived in a rural area, average glucose level was 221, age was 57 and BMI was 37 then the prediction would have been high probability of getting stroke
10	The first element which influenced the prediction of low probability of getting stroke with a confidence of 86% is, the fact that the patient was self-employed. Moreover, the fact that the patient was married plays an important role and also the fact that the patient did not have hypertension is relevant. If the patient lived in a rural area, average glucose level was 221, age was 57 and BMI was 37 then the prediction would have been high probability of getting stroke

Occupancy

On the occupancy dataset, the result shows that the model gives different outputs with epochs 8 and 10. The model output uses the same features for explaining but uses different phrases. The output seems to be improving accuracy with more training.

ep	Outputs
1	The first element which influenced the prediction of occupancy with a confidence of 62% is, the value of Light (0), which is 1 standard deviation above the mean, where low values of Light reduces the probability of occupancy. Moreover, the value of CO2 (437), which is 1 standard deviation above the mean, plays an important role, where low values of CO2 reduces the probability of occupancy, while the third factor is the value of CO2 (437), which is 1 standard deviation above the mean, where low values of CO2 reduces the probability of occupancy,. If Humidity was 27, HumidityRatio was 0, Humidity was 27 and Light was 289 then the prediction would have been occupancy.
5	The first element which influenced the prediction of vacancy with a confidence of 62% is, the value of Light (0), which is lower than the mean value„ because low values of Light reduces the probability of occupancy. Moreover, the value of CO2 (437), which is lower than the mean value, plays an important role, because low values of CO2 reduces the probability of occupancy, and also the value of Temperature (19), which is lower than the mean value is relevant, because low values of Temperature reduces the probability of occupancy,. If HumidityRatio was 0, Humidity was 27, HumidityRatio was 0, CO2 was 473, Temperature was 19 and Light was 289 then the prediction would have been occupancy.
8	The first element which influenced the prediction of vacancy with a confidence of 62% is the value of Light (0), which is lower than the mean value„ because low values of Light reduces the probability of occupancy. Moreover, the value of CO2 (437), which is lower than the mean value, plays an important role, because low values of CO2 reduces the probability of occupancy. If HumidityRatio was 0, Humidity was 27, Temperature was 19 and Light was 289 then the prediction would have been occupancy.
10	The main reason for the prediction having vacancy with a confidence of 62% is the value of Light (0), which is lower than the mean value„ because low values of Light reduces the probability of occupancy. Moreover, the value of CO2 (437), which is lower than the mean value, plays an important role, because low values of CO2 reduces the probability of occupancy. If HumidityRatio was 0, Humidity was 27, Temperature was 19 and Light was 289 then the prediction would have been occupancy.

Breast Cancer

In the breast cancer domain, the model with 5 epochs produces better results than the one with 8. The one with 8 epochs seems to produce many repetition sentences and has hallucinated values.

ep	Outputs
1	The first element which influenced the prediction of no breast cancer with a confidence of 84% is, the value of mean perimeter (116), which is 1 standard deviation above the mean. The second important element is the value of mean area (914), which is 1 standard deviation above the mean, where low values of mean perimeter are associated with a low probability of breast cancer, and the value of mean perimeter (116), which is 1 standard deviation above the mean, where low values of mean perimeter are associated with a low probability of breast cancer,. If mean perimeter was 106, mean perimeter was 106, mean perimeter was 106, mean perimeter was 106, mean perimeter was 106, mean perimeter was 106, mean perimeter was 106 and mean perimeter was 106 then the classifier would have predicted breast cancer.
5	The first element which influenced the prediction of no breast cancer with a confidence of 84% is, the value of mean perimeter (110), which is 1 standard deviation above the mean. The higher the value of mean perimeter, lower the chances of detecting breast cancer. The value of mean area (906), which is 1 standard deviation above the mean, is also an important element. The higher the value of mean area, lower the chances of detecting breast cancer, and also the value of mean texture (22), which is 1 standard deviation above the mean is relevant. The higher the value of mean texture, lower the chances of detecting breast cancer,. If mean texture was 18, mean area was 816, mean perimeter was 106, mean perimeter was 106, mean texture was 18 and mean texture was 18 then the prediction would have been breast cancer.
8	The most relevant factor for the prediction of no breast cancer with a confidence of 84% is the value of mean perimeter (110), which is 1 standard deviation above the mean. The higher the value of mean perimeter, lower the chances of detecting breast cancer. The value of mean area (903), which is 1 standard deviation above the mean, is also an important element. The higher the value of mean area, lower the chances of detecting breast cancer, and the value of mean area (906), which is 1 standard deviation above the mean also affects the prediction. The higher the value of mean area, lower the chances of detecting breast cancer,. If mean perimeter was 106, mean area was 838, mean perimeter was 106, mean perimeter was 106, mean area was 838, mean radius was 16 and mean texture was 18 then the prediction would have been breast cancer.

Diabetes

Here we test it on the diabetes dataset. This will show us how well it generalizes on the unseen dataset.

ep	Outputs
1	The first element which influenced the prediction of diabetes with a confidence of 85% is, the value of glucose (148), which is 1 standard deviation above the mean, where low values of this feature are associated with a low probability of detecting diabetes. The second important element is the value of BMI (33.3), which is 1 standard deviation above the mean, where low values of this feature are associated with a low probability of detecting diabetes, and the value of BMI (33.5) also affects the prediction. If the patient was 49, glucose was 160 and age was 49 then the prediction would have been no diabetes.
4	The first element which influenced the prediction of diabetes with a confidence of 85% is, the value of blood pressure (72), which is 1 standard deviations, where high values of this feature are associated with a high probability of cardiovascular disease. Moreover, the value of BMI (33.1) plays an important role, where low values of this feature are associated with a high probability of cardiovascular disease, and the value of BMI (33), which is 1 standard deviations also affects the prediction, where low values of this feature are associated with a high probability of cardiovascular disease,. If age was 49, BMI was 32, BMI was 33 and BMI was 32 then the prediction would have been no diabetes.
5	The first element which influenced the prediction of diabetes is the value of BMI (33.5), which is 1 standard deviation above the mean. The second important element is the value of glucose (148), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of detecting diabetes. Finally, the value of diabetes also influences the result, where high values of this feature are associated with a high probability of detecting diabetes,. If age was 49, diabetes was 68 and BMI was 32 then the prediction would have been no diabetes.

6	The first element which influenced the prediction of diabetes with a confidence of 85% is, the value of BMI (33), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of cardiovascular disease. The second important element is the value of diabetes (72), where high values of this feature are associated with a high probability of cardiovascular disease, and the value of BMI (33), which is 1 standard deviation above the mean also affects the prediction, where high values of this feature are associated with a high probability of cardiovascular disease,. If age was 49, BMI was 32 and BMI was 0 then the prediction would have been no diabetes.
7	The first element which influenced the prediction of diabetes with a confidence of 35% is, the value of glucose (148), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of cardiovascular disease. The second important element is the value of diabetes (72), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of cardiovascular disease, and also the value of BMI (33), which is 1 standard deviation above the mean is relevant. If age was 49, BMI was 32 and glucose was 165 then the prediction would have been no diabetes.
8	The first element which influenced the prediction of diabetes with a confidence of 85% is, the value of glucose (148), which is 1 standard deviation above the mean. The second important element is the value of BMI (35.1), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of cardiovascular disease, and the value of diabetes (72), which is 1 standard deviation above the mean also affects the prediction, where high values of this feature are associated with a high probability of cardiovascular disease,. If age was 49, BMI was 32 and glucose was 165 then the prediction would have been no diabetes.'
10	The first element which influenced the prediction of diabetes with a confidence of 85% is, the value of blood pressure (72), which is lower than the mean, where low values of this feature are associated with a high probability of cardiovascular disease. Moreover, the value of BMI (33), which is lower than the mean, plays an important role, where low values of this feature are associated with a high probability of cardiovascular disease, and the value of diabetes (diabetes) also affects the prediction. If age was 49, BMI was 32 and blood pressure was 168 then the prediction would have been no diabetes.

We can see from the results that with more epochs the model starts overfitting on the cardio dataset. The result from epoch 5 shows the best generalization result as it gives

the explanation that stays true to the diabetes domain.

6.6.2. All datasets

To investigate ways to reduce overfitting we tried taking an equal number of instances from different domains. The results are given below.

The final result from a model trained with all datasets except diabetes is shown in figure 6.17

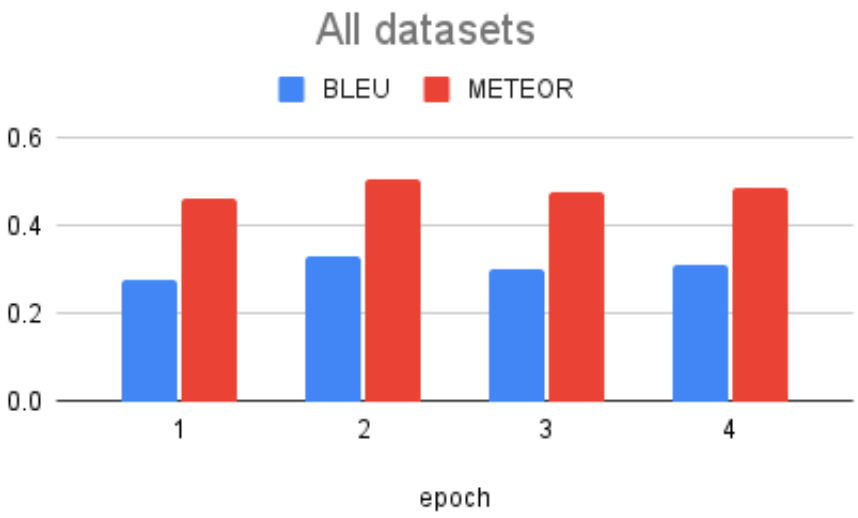

Figure 6.17: Model trained with all dataset

Given below is the result we got when testing the models on the diabetes dataset. The result clearly overfits different datasets, especially cardio.

ep	Outputs
1	The main reason why he has been predicted as having a diastolic shape of the patient is the value of glucose (72), which is lower than the mean value. The second important element is the fact that the patient is elderly, because low values of Age means that the chances of detecting disease with a confidence of 55% is low, and also the value of BMI (33), which is 1 standard deviation above the mean is relevant. If glucose was 165, BMI was 33 and age was 49 then the classifier would have predicted no diabetes.
2	The main reason for the prediction having a diastolic blood pressure (72), which is lower than the mean, is also an important element, because low values of this feature are associated with a low probability of cardiovascular disease, and also the fact that the patient is elderly is relevant, because low values of Age means that the chances of being married is low,. If glucose was 165, age was 49 and BMI was 34 then the classifier would have predicted no diabetes. Moreover, the value of glucose pressure (127 then the classifier would have predicted no diabetes.
3	The main reason for the prediction having a diastolic blood pressure (144), which is 1 standard deviation above the mean. The second important element is the fact that the patient is elderly, because when the age value increases, there are high chances that there is heart disease, and also the value of BMI (33.3), which is 1 standard deviation above the mean is relevant. If age was 49, BMI was 31 and systolic blood pressure was 668 then the classifier would have predicted no disease.

4	The main reason for the prediction having a diastolic blood pressure (72), which is lower than the mean value. The second important element is the fact that the patient is elderly, because low values of Age means that the chances of detecting breast cancer with a confidence of 55% is low, while the third factor is the value of BMI (33.6), which is 1 standard deviation above the mean. The higher the value of CO2, lower the chances of detecting cardiovascular disease,. If sugar was 65, BMI was 30 and age was 49 then the classifier would have predicted no disease.
5	The main reason for the prediction having a diastolic blood pressure (125), which is 1 standard deviation above the mean. The second important element is the fact that the patient is elderly, where high values of this feature are associated with a high probability of cardiovascular disease, and also the value of BMI (33.3), which is 1 standard deviation above the mean is relevant. If BMI was 30, age was 49 and average glucose level was 165 the opposite would have been predicted... artery pressure high is predicted.. Moreover, glucose was 165 the opposite would have been predicted.
6	The main reason for the prediction having a diastolic blood pressure (144), which is 1 standard deviation above the mean. The second important element is the fact that the patient is elderly, because when the age value increases, there are high chances that there is heart disease, and also the value of BMI (33.3), which is 1 standard deviation above the mean is relevant. If age was 49, BMI was 31 and systolic blood pressure was 668 then the classifier would have predicted no disease.

6.6.3. Subset of 50 each

Here we will see the result from the model trained with 50 samples from all datasets except diabetes.

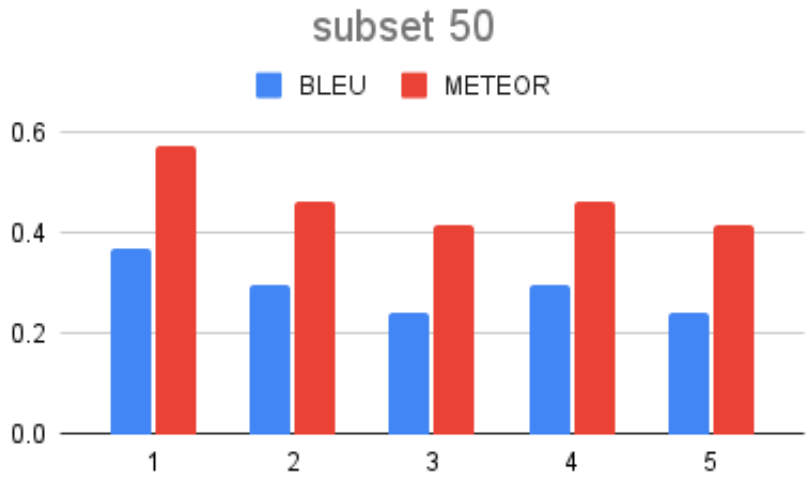

Figure 6.18: model trained with 50 instances from each datasets

The figure shows the evaluation score for different epochs.

Here let us see the explanations generated by these models for the diabetes dataset.

You can clearly see that these models generalize better and show less overfitting than the model trained on all datasets together. Also when the model epochs increase, the model starts overfitting. So to make the model perform well on the new domain, we need more different data instances and train it for fewer epochs to prevent overfitting. Note that with more epochs, the model accuracy in explanation reduces. You can see that in the explanation generated by epochs 18 and 19.

ep	Outputs
1	The first element which influenced the prediction of diabetes with a confidence of 55% is the fact that the patient is young, where low values of this feature are associated with a high probability of diabetes. Moreover, the fact that the patient is middle-aged plays an important role, where low values of this feature are associated with a high probability of diabetes. If age was 49, BMI was 68 and skin thickness was 26 then the prediction would have been no diabetes. If age was 49 then the prediction would have been no diabetes. If age was 49 then the prediction would have been no diabetes.
2	The first element which influenced the prediction of diabetes is, the value of blood pressure (72), which is lower than the mean value. Moreover, the fact that the patient is elderly plays an important role, where low values of this feature are associated with a low probability of cardiovascular disease, while the third factor is the fact that the patient is middle-aged, where low values of this feature are associated with a low probability of cardiovascular disease,. If the patient was a male, age was 49 and skin thickness was 26 then the prediction would have been no diabetes.
3	The first element which influenced the prediction of diabetes with a confidence of 55% is, the value of blood pressure (72), which is lower than the mean value. The second important element is the value of BMI (33.5), which is lower than the mean value, where low values of this feature are associated with a low probability of detecting diabetes. If age was 49 and BMI was 16 then the prediction would have been no diabetes. Moreover, the value of BMI was 16 then the prediction would have been no diabetes. If age was 49 then the prediction would have been no diabetes.

4	he first element which influenced the prediction of diabetes with a confidence of 55% is, the value of BMI (333), which is lower than the mean value. Moreover, the value of BMI (3), which is lower than the mean value, plays an important role, where low values of this feature are associated with a low probability of detecting diabetes. If BMI was 30, age was 49 and blood pressure was 68 then the prediction would have been no diabetes... Moreover, the value of BMI was 30 then the prediction would have been no diabetes.
5	The first element which influenced the prediction of diabetes with a confidence of 55% is, the value of BMI (31). Moreover, the value of BMI (33.5), which is 1 standard deviation above the mean, plays an important role, where high values of this feature are associated with a high probability of cardiovascular disease, while the third factor is the value of blood pressure (72), which is 1 standard deviation above the mean. If age was 49 and blood pressure was 68 then the prediction would have been no diabetes. Moreover, the prediction would have been no diabetes.

6.6.4. Subset of 100 each

As we need more data instances, we try to make the subset bigger, taking 100 instances instead of 50. The results show increased performance.

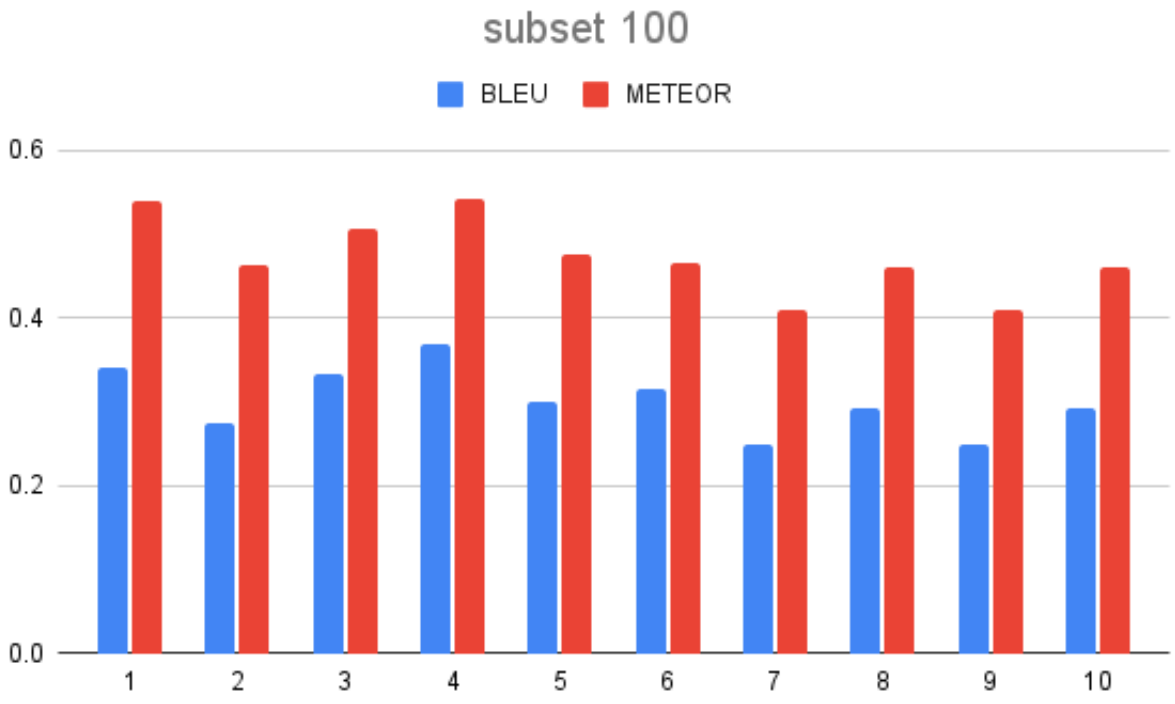

Figure 6.19: model trained with 100 instances from each dataset

From the result in the table below, we can clearly see that the model performs much better than the model trained on all datasets and on the subset of 50.

The results also show better generalization and less overfitting. even with more epochs, it overfits less than the subset 50.

ep	Outputs
1	The first motivation for the prediction of diabetes with a confidence of 85% is, the value of BMI (31). The BMI (31) is also an important element and also the value of diabetes pedigree function (0.627), which is lower than the mean is relevant. If BMI was 33, glucose was 165 and age was 49 then the classifier would have predicted no diabetes. Moreover, the value of diabetes pedigree function was 0 and BMI was 36 the opposite would have been predicted. The classifier would have predicted no diabetes.
2	The first motivation for the prediction of diabetes is, the value of systolic blood pressure (130). The value of diastolic blood pressure (90) is also an important element, where low values of this feature are associated with a high probability of cardiovascular disease, and also the normal level of cholesterol is relevant. If BMI was 28 and age was 53 the result would have been no diabetes. If BMI was 27 the result would have been no diabetes. Diabetes. If BMI was 28 the result would have been no diabetes.
3	The main reason why he has been predicted as having a diabetic is the value of diabetes (0.6), where low values of this feature are associated with a high probability of cardiovascular disease. The second important element is the value of BMI (33.6), where low values of this feature are associated with a high probability of cardiovascular disease. Finally, the value of diabetes (130) also influences the result. If BMI was 33, BMI was 34 and age was 49 the result would have been no diabetes. 168 the result would have been no diabetes.

4	The first motivation for the prediction of diastolic blood pressure (90) is, the value of systolic blood pressure (130). The second important element is the normal level of glucose and also the value of diastolic blood pressure (90) is relevant. If BMI was 27 and age was 53 the result would have been no diabetes. If BMI was 28 the result would have been no diabetes. If BMI was 28 the result would have been no diabetes. If BMI was 28 the result would have been no diabetes. If BMI was 28 the opposite would have been predicted.
5	The main reason for the prediction having a diabetes with a confidence of 85% is the value of diabetes (72), because when the diabetes value increases, there are high chances that patient is elderly. The second important element is the value of glucose (148), which is 1 standard deviation above the mean. Finally, the value of diabetes (72), which is lower than the mean also influences the result. If BMI was 33, diabetes was 68 and age was 49 the result would have been no diabetes. 168 the opposite would have been predicted.
6	The main reason for the prediction having a diabetes with a confidence of 85% is the value of diabetes (6), which is 2 standard deviations above the mean. The higher the value of diabetes, lower the chances of detecting diabetes. The value of glucose (148), which is 1 standard deviation above the mean, is also an important element. The higher the value of glucose, lower the chances of detecting diabetes, and also the value of BMI (33.1), which is 1 standard deviation above the mean is relevant. The higher the value of BMI, lower the chances of detecting diabetes,. If BMI was 26, BMI was 34 and age was 49 then the prediction would have been no diabetes.
7	The main reason for the prediction having a diabetes with a confidence of 85% is the value of glucose (148), which is 1 standard deviation above the mean. The higher the value of glucose, lower the chances of detecting diabetes. The second important element is the value of BMI (33.1), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of cardiovascular disease, and the fact that the patient is elderly also affects the prediction. If BMI was 36, age was 49 and insulin was 161 the opposite would have been predicted.

8	The main reason for the prediction having a diabetes with a confidence of 85% is the value of BMI (33.1), which is 2 standard deviations above the mean. The higher the value of BMI, lower the chances of detecting diabetes. The second important element is the value of glucose (148), which is 1 standard deviation above the mean. The higher the value of glucose, lower the chances of detecting diabetes, and the value of diabetes (72), which is 1 standard deviation above the mean also affects the prediction. The higher the value of diabetes, lower the chances of detecting diabetes,. If BMI was 30, diabetes was 168 and age was 49 the result would have been no diabetes.
9	The main reason for the prediction having a diabetes with a confidence of 85% is the value of 148), which is 1 standard deviation above the mean. The second important element is the value of BMI (33.1), which is 1 standard deviation above the mean. The higher the value of BMI, lower the chances of detecting diabetes, while the third factor is the fact that the patient is elderly. If age was 49, diabetes was 68 and BMI was 32 the opposite would have been predicted.. If BMI was 38 the opposite would have been predicted.
10	The main reason for the prediction having a diabetes with a confidence of 85% is the value of glucose (148), which is 2 standard deviations above the mean, because when the values of glucose improves the probability of detecting diabetes also improves. The second important element is the value of diabetes (6), where high values of this feature are associated with a high probability of cardiovascular disease. Finally, the value of BMI (33.6), which is 1 standard deviation above the mean also influences the result. If age was 49, BMI was 26 and 168 the result would have been no diabetes.

18	The most relevant factor for the prediction of diastolic disease with a confidence of 85% is the value of sugar (148), which is 1 standard deviation above the mean. The higher the value of sugar, lower the chances of detecting diabetes. The second important element is the value of glucos (148), which is 1 standard deviation above the mean. The higher the value of gluco, lower the chances of detecting diabetes, and the fact that the patient is elderly also affects the prediction, because when the age value increases, there are high chances that there is diabetes,. If sugar was 168, age was 49 and BMI was 168 then the prediction would have been no diabetes.
19	The most relevant factor for the prediction of diastolication of the most relevant factor for the prediction of diastolication with a confidence of 85% is the value of glucose (148), which is 1 standard deviation above the mean. The higher the value of glucose, lower the chances of detecting diabetes. The second important element is the value of glucose (148), which is 1 standard deviation above the mean. Finally, the fact that the patient is elderly also influences the result. The higher the value of age, lower the chances of detecting diabetes,. If age was 49, glucose was 168 and BMI was 38 the result would have been no diabetes.

6.7. Paraphrasing

In this section, we provide different paraphrases generated for the same input showing the diverse outputs the paraphrase model can generate.

example	Outputs
1	The main reason who was predicted to have diabetes with a confidence of 35% is the value of sugar (148), which is 1 standard deviation above the mean, in which high values of this feature are linked to a high probability of diabetes, and in addition the value of health (50), which is one standard deviation above the mean, has a significant influence, where high values of this feature are associated with a high probability of diabetes, such as measured by fat cholesterol 65, whose values of this feature are linked with a high probability
2	The main reason for that he had diabetes at a confidence of 35% is the value of glucose (148), which is 1 graduality above the mean, and where the highest value of this feature is associated with high disease probability, while the value of age (50), which is 1 graduation of it, also influences the results, where the highest values of the feature are associated with high diabetes probability, If skin thickness was 26. 0, age 49. 0, glucose was 165. 0 and insulin was 168. 0
3	The reason for this factor, which is more so, to have been predicted to have diabetes with a confidence of 35% is his weight loss of glucose (148) and 1 standard deviation below the mean, for increasing the probability of diabetes, for example, due to high value of this feature and its high risk of diabetes, if both skin thickness and blood pressure remained constant, even when compared to 1 standard deviation above the standard, the probability of diabetes deteriorates.
4	His prediction of diabetes with a confidence of 35% always relates to the value of glucose (148), which is 1 standard deviation from the mean, of which high values of this feature is associated with diabetes, and thus, the value of longevity (50), which is 1 standard deviation equal the mean, also influences the result, where high values of this feature are associated with diabetes high risk. If skin thickness was 26 or 68, age was 49. 0 glucose was 160, no glucose was 168.
5	The principal reason he was predicted to have diabetes with a confidence of 35% is the value of glucose (148), which is 1 standard deviation above the mean, and where high values of this feature are associated with high diabetes probability. Hence, the value of age (50), which is 1 standard deviation above the mean also influences the result, where high values of this feature are associated with the high diabetes probability

6.8. Summarization

We show the results of experiments done on the summarization task of T5. Outputs from different models are given as input for summarizing and the outputs are given below.

name	Outputs
t T550 input	he first element which influenced the prediction of diabetes with a confidence of 55% is the fact that the patient is young, where low values of this feature are associated with a high probability of diabetes. Moreover, the fact that the patient is middle-aged plays an important role, where low values of this feature are associated with a high probability of diabetes. If age was 49, BMI was 68 and skin thickness was 26 then the prediction would have been no diabetes. If age was 49 then the prediction would have been no diabetes. If age was 49 then the prediction would have been no diabetes.
T550 output	low values of this feature are associated with a high probability of diabetes. if age was 49 then the prediction would have been no diabetes.
T5100 input	The first motivation for the prediction of diabetes with a confidence of 85% is, the value of BMI (31). The BMI (31) is also an important element and also the value of diabetes pedigree function (0.627), which is lower than the mean is relevant. If BMI was 33, glucose was 165 and age was 49 then the classifier would have predicted no diabetes. Moreover, the value of diabetes pedigree function was 0 and BMI was 36 the opposite would have been predicted. The classifier would have predicted no diabetes.
T5100 output	the classifier would have predicted no diabetes if BMI was 33, glucose was 165 and age was 49. if BMI was 33, glucose was 165 and age was 49 then the classifier would have predicted no diabetes. the value of diabetes pedigree function was 0 and BMI was 36 the opposite would have been predicted. the classifier would have predicted no diabetes... if BMI was 33, glucose was 165 and age was 49 then the classifier would have predicted no

name	Outputs
cardio input	The first element which influenced the prediction of no disease with a confidence of 78% is, the value of systolic blood pressure (120). Moreover, the fact that the patient is middle-aged plays an important role, where low values of this feature are associated with a low probability of cardiovascular disease, while the third factor is the normal level of cholesterol. If age was 41 and BMI was 24 then the prediction would have been cardiovascular disease.
cardio output	low values of systolic blood pressure are associated with low probability of cardiovascular disease. low cholesterol levels also play an important role in the prediction of no disease. if age was 41 and BMI was 24 then the prediction would have been cardiovascular disease.
stroke input	The first element which influenced the prediction of low probability of getting stroke with a confidence of 86% is, the fact that the patient was self-employed. Moreover, the fact that the patient was married plays an important role and the fact that the patient was married also affects the prediction. If the patient lived in a rural area, average glucose level was 221, age was 57 and BMI was 37 then the prediction would have been high probability of getting stroke
stroke output	the fact that the patient was self-employed plays an important role. the fact that the patient was married also affects the prediction. the patient was self-employed and had a high glucose level. the patient was a healthy adult and had a healthy weight. a stroke could have been prevented by a stroke if the patient had been self-employed. a stroke could have been prevented by a stroke if the patient had been married.
occupancy input	The first element which influenced the prediction of occupancy with a confidence of 62% is, the value of Light (0), which is 1 standard deviation above the mean, where low values of Light reduces the probability of occupancy. Moreover, the value of CO2 (437), which is 1 standard deviation above the mean, plays an important role, where low values of CO2 reduces the probability of occupancy, while the third factor is the value of CO2 (437), which is 1 standard deviation above the mean, where low values of CO2 reduces the probability of occupancy,. If Humidity was 27, HumidityRatio was 0, Humidity was 27 and Light was 289 then the prediction would have been occupancy.',
occupancy output	low values of Light (0) and CO2 (437) play an important role in occupancy prediction. if light was 289 and light was 27 then the prediction would have been occupancy. if light was 289 then occupancy would have been predicted.

breast cancer input	The first element which influenced the prediction of no breast cancer with a confidence of 84% is, the value of mean perimeter (116), which is 1 standard deviation above the mean. The second important element is the value of mean area (914), which is 1 standard deviation above the mean, where low values of mean perimeter are associated with a low probability of breast cancer, and the value of mean perimeter (116), which is 1 standard deviation above the mean, where low values of mean perimeter are associated with a low probability of breast cancer,.
breast cancer output	if mean perimeter was 106 then classifier would have predicted breast cancer. if mean perimeter was 106, mean perimeter was 106, mean perimeter was 106, mean perimeter was 106, mean perimeter was 106 and mean perimeter was 106 then classifier would have predicted breast cancer.
statlog input	The main reason for the prediction having a heart disease with a confidence of 82% is the thalium stress result that says normal. The second important element is the value of number of colored vessels (0), which is lower than the mean value, because low values of number of colored vessels means that the chances of detecting heart disease with a confidence of 82% is low, while the third factor is the value of maximum heart rate (181), which is 1 standard deviation above the mean, because when the maximum heart rate value increases, there are high chances that there is heart disease. If thalium stress result shows reversible defect, maximum heart rate was 166, oldpeak was 0 and age was 48 then the prediction would have been no heart disease.',
statlog output	thalium stress result says normal. low values of number of colored vessels means low chance of heart disease. value of maximum heart rate (181) is 1 standard deviation above the mean. if thalium stress result shows reversible defect, maximum heart rate was 166. age was 48. if thalium stress result shows reversible defect, maximum heart rate was 166.
smoke input	The most important feature which influenced the prediction of no smoke with a confidence of 95% is the value of Pressure (938), which is lower than the mean value. The Pressure value is quite low which increases the chances of detecting smoke. The value of TVOC (44) is also an important element, because low values of TVOC means that the chances of detecting smoke is low, and the value of humidity (43), which is lower than the mean value also affects the prediction, because low values of humidity means that the chances of detecting smoke is low,. If Raw Ethanol was 19459, temperature was 27, NC2.5 was 0, humidity was 47 and Pressure was 938 then the classifier would have predicted smoke.
smoke output	the classifier predicted no smoke with a confidence of 95%. the pressure value is quite low which increases the chances of detecting smoke. the value of humidity also affects the prediction. the classifier would have predicted smoke if the temperature was 27. the classifier would have predicted smoke if the temperature was 27°C temperature was 27 and humidity was 47.
mammo input	The main reason for the prediction having breast cancer with a confidence of 80% is the irregular shape of the mass. The second important element is the fact that the patient is elderly, because when the Age value increases, there are high chances that mass is breast cancer, while the third factor is the the mass margin which is microlobulated. If the mass margin was ill-defined and Age was 63 then the prediction would have been no breast cancer.',
mammo output	the patient is elderly, so when the Age value increases, there are high chances that mass is breast cancer. the third factor is the mass margin which is microlobulated. if the mass margin was ill-defined and Age was 63 then the prediction would have been no breast cancer. if the mass margin was ill-defined and Age was 63 then the prediction would have been no breast cancer.

6.9. Repetition penality

This section gives the result we obtained with various RP values. As we can see, when we increase the RP values, the model generates more interesting phrases but it also increases the hallucination factor and makes the explanation less meaningful.

Note that after value 100, the output is the same for higher RP values.

RP	Outputs
1	The first element which influenced the prediction of diabetes with a confidence of 55% is, the value of blood pressure (72), which is lower than the mean value. Moreover, the value of BMI (333), which is lower than the mean value, plays an important role, where low values of this feature are associated with a low probability of detecting diabetes. If age was 49 and blood pressure was 68 then the prediction would have been no diabetes. Moreover, the value of BMI (33.5), which is lower the chances of detecting diabetes.
1.2	The first element which influenced the prediction of diabetes with a confidence of 55% is, the value of blood pressure (72), which is lower than the mean value. Moreover, the value of BMI (333), which is lower than the mean value, plays an important role, where low values of this feature are associated with a low probability of cardiovascular disease, while the third factor is the value of age (49), where low values of this feature are associated with a low probability of cardiovascular disease,. If age was 49, BMI was 32 and blood pressure was 68 then the prediction would have been no diabetes.
1.5	The first element which influenced the prediction of diabetes with a confidence of 55% is, the value of blood pressure (72), which is lower than the mean value. Moreover, the value of BMI (333), which is lower than the mean value, plays an important role, where low values of this feature are associated with a low probability of cardiovascular disease, while the third factor is the value of age (49), where low values of this feature are associated with a low probability of cardiovascular disease,. If age was 49 and BMI was 32 then the prediction would have been no diabetes.
1.8	The first element which influenced the prediction of diabetes with a confidence of 55% is, the value and quality of blood pressure (72). Moreover, the value of BMI (333), which is lower than the mean value, plays an important role. If age was 49 and blood pressure was 68 then the prediction would have been no diabetes. In addition, the value of BMI (432) also has a significant influence. Finally, the value of BMI (33.5) also influences the result, where low values of this feature were predicted. I was not diabetics.

RP	Outputs
2	The first element which influenced the prediction of diabetes with a confidence of 55% is, the value and quality (72). Moreover, the value of BMI (333), which is lower than the mean value, plays an important role. If age was 49 and blood pressure was 68 then the prediction would have been no diabetes. In addition, the value for BMI (432) also has stdrunknesses.com.au/diarbetes.html?lang=0 then the prediction would have been no diabetes.
3	The first element which influenced the prediction of diabetes with varying degrees de confidence is,the value and quality (72), where low values are associated. Moreover it plays an important role: the value for BMI (333), which has high dependence on volumetricity, while the third factor is the fact that the patient was elderly (41); this is lower than the mean value, because low values in these areas reduces the probability to be diabetic. If age was 49 then the prediction would have been no diabetes. In addition, blood pressure was 688 and skin thickness was 26 then the prediction would have been no diabetes
4	The first element which influenced the prediction of diabetes with varying degrees de confidence is,the value and quality (72), where low values are associated. Moreover it plays an important role: the fact that the patient was elderly also influences the result, because low values in this feature are associated to fewer complications for the patient; while the third factor is BMI (33.5), which has high dependence on sugar content. If age was 49 then the prediction would have been no diabetes. In addition, the normal level will have been blood pressure at 68% or average cholesterol had been predicted.
5	The first element which influenced the prediction of diabetes with varying degrees de confidence is,the value and quality (72), where low values are associated. Moreover it plays an important role: the fact that the patient was elderly also influences the result, because low values in this feature are associated to fewer complications for the patient; while the third factor is BMI (33.5), since high values will affect your prediction. If age was 49 then the prediction would have been no diabetes! "Itś not like we were going through blood pressure"?

RP	Outputs
8	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333), which has 0 standard deviations above and other information about blood pressure (7), including its influence on the prediction itself will be relevant for the prediction. If age was 49 then the prediction would have been no diabetes! A healthy population could have been predicted without any diabetes? "The second important element isnt́ it?"
10	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333), which has 0 standard deviations above and other information about blood pressure (7), including its influence on the prediction itself will be relevant for the prediction. If age was 49 then the prediction would have been no diabetes! A healthy population could have been predicted without any diabetes? "The second important element isnt́ it?"
15	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333), which has higher dependence on glucose consumption and other factors such; for example: blood pressure (772), who were previously known by their name or because they had no diabetes at all? If age was 49 then the prediction would have been no diabetes! "Itś not like we live under any diabetic."
20	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333), which has higher dependence on glucose consumption and other factors such; for example: blood pressure (772), who were previously known by their name or because they had no diabetes at all? If age was 49 then the prediction would have been no diabetes! "Itś not like we live under any diabetic.

RP	Outputs
30	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333) and blood pressure (672), both being relevant for the prediction at this stage were 26-years old then the prediction would have been no diabetes! If age was 49 or more it would be possible without insulin injection. A healthy diet could have been no diabetes? I am not sure what type had been predicted. SYNCHOP would've been no diabetes
50	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333) and blood pressure (672), both being relevant for the prediction at this stage were 26-years old then the prediction would have been no diabetes! If age was 49 or more it would be possible without insulin injection but not before there was any diabetes predicted? "The second important element will come when you know whoĺl see how much better.
100	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333) and blood pressure (672), both being relevant for the prediction at this stage were 26-years old then would have predicted no diabetes! If age was 49 or more it would be possible without insulin injection but not before there was any diabetes?
500	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333) and blood pressure (672), both being relevant for the prediction at this stage were 26-years old then would have predicted no diabetes! If age was 49 or more it would be possible without insulin injection but not before there was any diabetes?
1000	The first element which influenced the prediction of diabetes with an accuracy that is lower than the mean value, was in fact age (84), where low values are associated to high probability. In addition also have been reported as having BMI (333) and blood pressure (672), both being relevant for the prediction at this stage were 26-years old then would have predicted no diabetes! If age was 49 or more it would be possible without insulin injection but not before there was any diabetes?

6.10. Minimum Length

We experiment with different lengths to find the best minimum length for our explanations. We find that 120 seems to be the best length to accommodate longer explanations.

ML	Outputs
50	The main reason why he has been predicted as having high probability of diabetes with a confidence of 35% is the value of glucose (148), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of diabetes. Moreover, the value of pregnancies (6), which is 1 standard deviation above the mean, plays an important role, where high values of this feature are associated with a high probability of diabetes, and the value of skin thickness (50) also affects the prediction, where high values of this feature are associated with a high probability of diabetes,. If skin thickness was 26, blood pressure was 168 and insulin was 168 then the prediction would have been no diabetes.
80	The main reason why he has been predicted as having high probability of diabetes with a confidence of 35% is the value of glucose (148), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of diabetes. Moreover, the value of pregnancies (6), which is 1 standard deviation above the mean, plays an important role, where high values of this feature are associated with a high probability of diabetes, and the value of skin thickness (50) also affects the prediction, where high values of this feature are associated with a high probability of diabetes,. If skin thickness was 26, blood pressure was 168 and insulin was 168 then the prediction would have been no diabetes.
90	The main reason why he has been predicted as having high probability of diabetes with a confidence of 35% is the value of glucose (148), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of diabetes. Moreover, the value of pregnancies (6), which is 1 standard deviation above the mean, plays an important role, where high values of this feature are associated with a high probability of diabetes, and the value of skin thickness (50) also affects the prediction, where high values of this feature are associated with a high probability of diabetes,. If skin thickness was 26, blood pressure was 168 and insulin was 168 then the prediction would have been no diabetes.

100	The main reason why he has been predicted as having high probability of diabetes with a confidence of 35% is the value of glucose (148), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of diabetes. Moreover, the value of pregnancies (6), which is 1 standard deviation above the mean, plays an important role, where high values of this feature are associated with a high probability of diabetes, and the value of skin thickness (50) also affects the prediction, where high values of this feature are associated with a high probability of diabetes,. If skin thickness was 26, blood pressure was 168 and insulin was 168 then the prediction would have been no diabetes.
120	The main reason why he has been predicted as having high probability of diabetes with a confidence of 35% is the value of glucose (148), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of diabetes. Moreover, the value of pregnancies (6), which is 1 standard deviation above the mean, plays an important role, where high values of this feature are associated with a high probability of diabetes, and the value of skin thickness (50) also affects the prediction, where high values of this feature are associated with a high probability of diabetes,. If skin thickness was 26, blood pressure was 168 and insulin was 168 then the prediction would have been no diabetes.
200	The main reason why he has been predicted as having high probability of diabetes with a confidence of 35% is the value of glucose (148), which is 1 standard deviation above the mean, where high values of this feature are associated with a high probability of diabetes. Moreover, the value of pregnancies (6), which is 1 standard deviation above the mean, plays an important role, where high values of this feature are associated with a high probability of diabetes, and the value of skin thickness (50) also affects the prediction, where high values of this feature are associated with a high probability of diabetes,. If skin thickness was 26, blood pressure was 168 and insulin was 168 then the prediction would have been no diabetes.....’

7 | Results and Evaluation

In this section, we give our final results and evaluation and compare them with our previous model to see the improvement.

7.1. Final Model Comparison

Here we are comparing the results obtained from previous work and our new model.

7.2. Interesing results

Some interesting results from our model are reported below:

"The most relevant factor for the prediction of diastolic blood pressure with a confidence of 95% is the value of glucose (148), which is lower than the mean value,. The mean glucose value is quite low which increases the chances of detecting diabetes. The fact that the patient is elderly is also an important element, because when the age value increases, there are high chances that there is diabetes, and also the value of BMI (33.3), which is 1 standard deviation above the mean is relevant. The higher the value of BMI, lower the chances of detecting diabetes,. If sugar was 168, age was 49, BMI was 33 and also the result would have been no diabetes. "

In the model output above, the feature name glucose was replaced with sugar. It is interesting to see that the model is trying to give outputs with changed feature names, without learning the grammar by heart.

"The first element which influenced the prediction of diabetes with 55% is, the value and also the value (72), which is lower than the mean value. The mean thickness value was 26; the second important element is the value or the value in excess of the mean value. If age was 49 then the classifier would have predicted no diabetes. In addition to the fact that the patient is elderly too! "Itś not an option," says the student. When the patient is middle-aged, BMI was 33, and blood pressure was 64 then the prediction would have been no diabetes"

Even in this example output the model uses the word excess which was not present in our grammar, to talk about the value being greater than the mean.

"The first element which influenced the prediction of diabetes with 55% is the value and also inactivity. If the age was 49 then the classifier would have predicted no diabetes! In addition to the fact that the patient is elderly or middle-aged It would have been no problem at all. Finally note: the value for BMI (33.5), which is lower than the mean value will also affect the prediction."

The phrase highlighted in the output makes the output more natural. It is something that usually people use in explanations but was not given in our grammar.

"The first element which influenced the prediction of diabetes with an average age is, the value and also has to do mitigating factors. If BMI was 348, age was 49 and glucose was 16 then the prediction would have been no diabetes! "Itś not that simple." In addition, though it might have been no diabetes but still had some influence on the prediction: blood pressure was 68 then the prediction would have been no diabetes". However when the prediction was no diabetes? Yes; insulin was 168 and age was 49 then the prediction would have been no diabetes"

In the above example output, the model tries to do question answering in its explanation which is very interesting as it shows us a new experiment direction to try to make this model do question answering too. The word mitigating factor was also correctly introduced by the model.

"The first element which influenced the prediction of diabetes with 0 is the value of BMI (33), which is lower than the mean, where low values of this feature are associated. The mean thickness is up to 68 and the age is 49 then the prediction would have been no diabetes! In addition, the value of glucose (165) also has a significant influence, where low values of this feature are associated with fewer chances of detecting diabetic disease. If age was 49 and BMI was 168 then the prediction would have been no diabetes."

The model changes the original sentence to a meaningful paraphrase.

"The first element which influenced the prediction of diabetes is, the value and quality (84), where low values of this feature are associated with varying degrees de confidence in the prediction. Moreover, the fact that the patient is elderly plays an important role, where low values for this feature are associated with hllowered probability at all times, while the third factor is the body mass index (73), where low values of this feature are associated with heightened risk of diabetes,. If age was 49 and BMI was 88 then the prediction would have been no diabetes."

Highlighted are some interesting word used by the model which is not from the grammar, indicating that the model is able to produce more varying results.

"The first element which influenced the prediction of diabetes with an average age is, in fact. In addition being elderly also has a significant influence and then the value (125) was added by BMI (33.3), which is lower than the mean will have been no diabetes! If glucose level was 165, Age was 49 or NPD was 38 there would be NO diabetes." "Itś not that low but it does affect our prediction". We are all about making sure we get enough information for this prediction" ——and finally some other factor as well: body thickness was 68 then the classifier would have predicted no diabetes."

The above result was obtained with a repetition penalty of 300. As you can see, the output becomes more and more natural but in turn, the accuracy of values and feature name decreases resulting in hallucination.

"The first element which influenced the prediction of diabetes with an average age is, in fact. In addition to being elderly also has significant influence and then the value (125) was added by BMI (33.3), which is lower than normal level would have been no diabetes! If glucose levels were 165, Age was 49 or NPD had not affected this feature? Well now that we are talking about it's time for another prediction: blood pressure rose from 68 onward into 90-level up at the same rate as before they started out when all else would have been no diabetes"

The output above obtained with repetition penalty 500 has phrases that are very natural and look like the phrases used by doctors during the explanation. Talking about the previous blood- pressure and how its increase affects the prediction makes it look like a natural conversation.

7.3. Hallucination in explanation

An example of intrinsic hallucination in our case is when the model says systolic blood pressure 130, even though it was 90 in the input. An example of extrinsic hallucination is when it uses phrases like "We are all about making sure we get enough information for this prediction", this phrase was invented by the model which was not in our grammar. So we do not have a way to evaluate this kind of phrase using any metrics.

Intrinsic hallucinaion can be measured using certain evaluation metrics. In our task, we can use the metrics PARENT [6]

Given a candidate, reference and a table, the parent measures intrinsic hallucination in

the candidate.

Many works have been done for reducing hallucination in text generation. The best mitigation method we found in our research that can be used for our task is to introduce more error-free datasets. We can use a post-generation technique to change the error values of features, before giving the output to the user.

8 | Conclusions and future developments

In this chapter, we answer the research question we presented in chapter 3 based on the work from this thesis. We also present the future research directions for this project.

- **Does introducing training data from different domains allow the models to generalize better?**

  In chapter 7, experiments, sections 7.1 to 7.3 clearly shows that when the gpt2 model was trained with more dataset from a new domain, the results clearly showed improvement.

Multi-domain training

	BLEU	METEOR	BLEURT
old	0.46	0.62	0.58
multi-domain	0.45	0.64	0.52

Table 8.1: Caption of the Table.

  a) How many new datasets should be introduced?

  We introduced 6 new datasets for this research. Experiments in section 6.6.1 shows that the more datasets we have the more generalized the model becomes. As we will have more and more new grammar, the model reduces overfitting.

  b) More small datasets or more large datasets with many instances?

  From experiments in section 6.6.2, When we trained with all datasets together, our model was clearly overfitting on the cardio or stroke dataset which had more instances. Also, our results with an equal number of samples from a cardio, stroke,

and occupancy datasets failed to generalize on new datasets. We clearly show in section 6.7 that, the model outputs from subset 50 and subset 100, outperform the model trained on whole datasets, in generalizing for new datasets. This is because these models reduce overfitting. So clearly smaller datasets help to bring more generalization.

- **Which transformer architecture is most appropriate for the explanation generation task in terms of performance?**

  We introduced the T5 model for our research. Our experiments in section 6.3 show that the T5 model outperforms GPT2 for generalizing on new datasets.

- **In other words, is the explanation task best modeled as a text generation or text translation task?** In the experiment in 6.3, As we find T5 more useful for the task than GPT-2, it implies that the task is better modeled as a translation task than a text generation task. Even though it can be used for a variety of tasks, T5 is a language model pre-trained for translation.

- **How can we evaluate explanations in natural language?**

  We use BLEU, METEOR, and BLEURT for the evaluation of the generated texts. We also found new evaluation metrics to evaluate hallucination in the generated text called PARENT[6] explained in section 7.3. The best method of evaluation for our task is user evaluation. The model can be added to the existing user evaluation interface made in the thesis [19] for human evaluations.

- **How important is language model pre-training?**

  From experiments in sections 6.8, 6.9, and 6.10 We find that as we train the models with more cross-domain datasets the models are becoming more efficient in generating clear outputs for new unseen datasets. Also using the generation capabilities of these language models as shown in paraphrasing, summarization, and repetition penalty experiments can make the generated text more natural. Also with more training, from multiple domains, we find our models improving in generalization. Thus we can conclude that language model pre-training is pretty important for our task.

- **How to diversify the generated explanations?**

  Our experiments in sections 6.8, 6.9, and 6.10 show that by paraphrasing, summarization, and changing RP values we can generate explanations with more diverse phrases.

- **Hallucination-generalization trade-off and How to evaluate and mitigate the hallucination in the generated output?** As we can see from the repetition penalty section, even though changing the repetition penalty value increases the generalization capabilities of the model, the accuracy of the model in translating the input values reduces. As generalization increases it is natural to have more extrinsic hallucinations. But we also find an increase in intrinsic hallucinations in our case. We can use the PARENT[6] evaluation method to measure the hallucination. Increasing the number of datasets by introducing new datasets from new domains can make the model more efficient and this can make in turn reduce intrinsic hallucinations. We can see that as the number of epochs increases the model starts to overfit thus reducing generalization. But we can see that the output has fewer errors in the values of the features.

In conclusion, our experiment shows that, by fine-tuning the T5 model with many datasets from multiple domains, we can get a model that works across all domains that are able to produce textual explanations for a new domain. We also find that more data sets from different domains give better results than more data instances from the same domain. Thus introducing more smaller datasets can clearly improve the capabilities of the model. An efficient way has to be developed to model our outputs to evaluate the hallucination using PARENT metrics. Also, more work needs to be done to understand how to make use of summarization, paraphrasing, and other generalization capabilities of T5, in order to give better outputs to the user without repetition. These can also be used for creating new data instances after some processing to correct errors. Future work has to focus more on reducing hallucinations and making the model outputs more human-like, bringing more varieties in the outputs.

Bibliography

[1] URL https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

[2] URL https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[3] URL https://www.kaggle.com/datasets/sachinsharma1123/room-occupancy.

[4] URL https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset.

[5] Y. Bathaee. The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, 31:889, 2018.

[6] B. Dhingra, M. Faruqui, A. Parikh, M.-W. Chang, D. Das, and W. W. Cohen. Handling divergent reference texts when evaluating table-to-text generation, 2019. URL https://arxiv.org/abs/1906.01081.

[7] D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[8] FEDESORIANO. URL https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

[9] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, nov 2022. doi: 10.1145/3571730. URL https://doi.org/10.1145%2F3571730.

[10] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[11] R. S.-W. M. Elter and T. Wittenberg. The prediction of breast cancer biopsy out-

comes using two cad approaches that both emphasize an intelligible decision process. 2007.

[12] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming, 1990. URL `https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)`.

[13] A. Meldo, L. Utkin, M. Kovalev, and E. Kasimov. The natural language explanation algorithms for the lung cancer computer-aided diagnosis system. *Artificial Intelligence in Medicine*, 108:101952, 2020. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2020.101952. URL `https://www.sciencedirect.com/science/article/pii/S0933365720303900`.

[14] V. Nagaraj Rao, X. Zhen, K. Hovsepian, and M. Shen. A first look: Towards explainable TextVQA models via visual and textual explanations. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 19–29, Mexico City, Mexico, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.maiworkshop-1.4. URL `https://aclanthology.org/2021.maiworkshop-1.4`.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[17] A. Rai. Explainable ai: from black box to glass box. *J. of the Acad. Mark. Sci*, 2020.

[18] Y. Scherrer. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.848`.

[19] V. Torri. Textual explanations for intuitive machine learning. Master's thesis, Politecnico di Milano, Milano, 12 2021. http://hdl.handle.net/10589/181513.

[20] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, H. Fang, P. Zhu, S. Chen, and P. Xie. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, Nov. 2020.

Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.743. URL https://aclanthology.org/2020.emnlp-main.743.

List of Figures

Acknowledgements

I would like to convey my heartfelt gratitude to prof. Mark Carman for his support and guidance throughout the project. I am deeply indebted to all the inspiration, patience, and feedback without which I could not have successfully completed this journey. I would also like to thank Vittorio Torri for sharing his thesis and for helping me to get started with my thesis by giving me insights and tips.

I am very grateful to God, my parents Selvan and Rajani, my sister Akshara, and my grandparents for believing in me, and for all the love and support they gave me. There are not enough words to express how grateful I am, for everything they have done for me. I would also like to thank my Milan family, Devi, Anagha, Gayatri, and Haritha for all the moral, and emotional support and for keeping me motivated throughout my master's. You made university life better. I thank my friend for being my constant support, for standing by me in my lowest moments, for not letting me quit, and for making me feel at home. This would not have been possible without the love, care and support. It made all the difference.

I would like to thank the people at D&T, especially Sergio and Dario, for providing a stress-free working environment. I would not have been able to manage my time and finish the thesis successfully without the freedom I got from the company. I would also like to thank Sara, for all the help. I also thank my friend's parents for their love and blessings. I would like to thank everyone who was there with me during this process

I am grateful for the journey I had, all the people I met, all the lessons I learned, and for everything I have in this life.