Evaluation of LIME and SHAP in Explaining Automatic ICD-10 Classifications of Swedish Gastrointestinal Discharge Summaries

Alexander Dolk[1], Hjalmar Davidsen[1], Hercules Dalianis[1,2], Thomas Vakili[1]
[1]Department of Computer and Systems Sciences (DSV) Stockholm University, Kista, Sweden
[2]Norwegian Centre for E-health Research, Tromsø, Norway
alexander.dolk@hotmail.com, hjalmar.davidsen@gmail.com,
hercules.dalianis@ehealthresearch.no, thomas.vakili@dsv.su.se

Abstract
A computer-assisted coding tool could alleviate the burden on medical staff to assign ICD diagnosis codes to discharge summaries by utilising deep learning models to generate recommendations. However, the opaque nature of deep learning models makes it hard for humans to trust them. In this study, the explainable AI models LIME and SHAP have been applied to the clinical language model SweDeClin-BERT to explain ICD-10 codes assigned to Swedish gastrointestinal discharge summaries. The explanations have been evaluated by eight medical experts, showing a statistically higher significant difference in explainable performance for SHAP compared to LIME.

Keywords
ICD-10 diagnosis code, natural language processing, eXplainable AI, multi-label text classification.

1 INTRODUCTION

The International Classification of Diseases (ICD) has been used globally for over a century to classify information in patient records [1]. Using the ICD coding system, reported conditions in patient records are converted into medical codes. The coded patient records are then used for administrative and research purposes. The ICD coding system has been revised multiple times. Currently, the tenth version (ICD-10) is the most widely used edition.

The ICD framework is important as it is a common way of recording diseases, enabling health practitioners within and between countries to share their data. The ICD promotes the compilation and storage of medical data for decision-making and analysis. Currently, the ICD is used by all member states of the World Health Organization and has been translated into 43 languages [2].

Human coders are prone to making errors when assigning ICD-10 codes. For example, one study [3] found that 20 percent of the main diagnoses in Swedish discharge summaries were incorrectly coded.

A computer-assisted coding (CAC) tool for ICD-10 coding that utilises artificial intelligence (AI) can give recommendations to physicians on possible ICD-10 codes for a discharge summary. In addition, it can also validate already assigned ICD-10 diagnosis codes. The use of such tools has the potential to increase the efficiency of the health care system.

The development of a CAC-tool for ICD-10 coding is highly needed in the medical field. This is part of the ClinCode project, at the Norwegian Centre for E-health Research [4].

The use of artificial intelligence (AI) and machine learning has in recent years become widespread. Novel strategies like deep learning (DL) models have demonstrated great results for a multitude of regression and natural language processing (NLP) tasks [5][6]. Nonetheless, DL models are opaque in nature. It is often impossible for humans to understand why DL models make particular predictions. This is an issue as it makes it hard for humans to trust the predictions of DL models [7]. To remedy this problem, the field of explainable artificial intelligence (XAI) has recently emerged [8]. The purpose of XAI is to provide methods that can explain the prediction of AI models.

In this article, the XAI models Local Interpretable Model-agnostic Explanations (LIME) [9], and SHapley Additive exPlanations (SHAP) [10] have been applied post hoc to the classification model **Swe**dish **De**-identified **Clin**ical **BERT** (SweDeClin-BERT) [11], to explain ICD-10 classifications of Swedish gastrointestinal discharge summaries. SweDeClin-BERT is a derivation of the model KB-BERT, a **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) model [12] developed by the National Library of Sweden (KB). The explanations by LIME and SHAP have then been evaluated by medical doctors and ICD coding experts through a questionnaire, resulting in a comparison of the model's explainable performance. Specifically, the models have been compared for the factors of user trust, explanation satisfaction and perceived usability.

While there have been previous evaluations of LIME and SHAP with medical data, only one peer-reviewed study has been found which applies SHAP with Swedish medical data [13]. The domain dependence of data when evaluating

The 18th Scandinavian Conference on Health informatics, Tromsø, Norway, August 22-24, 2022. Organized by UiT The Arctic University of Norway. Conference Proceedings published by Linköping University Electronic Press at https://doi.org/10.3384/ecp187.

AI models motivates the need for this article, as it uses medical data labelled with ICD-10 codes in Swedish. To the extent of our knowledge, this paper is the first that evaluates LIME and SHAP on Swedish medical data labelled with ICD-10 codes. From a Human-Computer Interaction (HCI) perspective, it is also valuable to get feedback from respondents of the population that will be future users of a CAC-tool for ICD-10 coding. It is valuable as the performance of XAI models is highly subjective depending on the user group.

2 RELATED RESEARCH

This article builds upon previous research [14], which has evaluated the deep learning model KB-BERT against a range of baseline models for the task of multi-labelling Swedish gastrointestinal discharge summaries with ICD-codes. The results of that article showed that a fine-tuned version of KB-BERT achieved an F_1-micro of 0.80 and F_1-macro of 0.58 on grouped ICD-10 codes. However, when tested on the full 263 ICD codes, the KB-BERT model underperformed against the baseline models. In the article, it was recommended to further study the possibility of including explainability mechanisms in a CAC-tool for ICD-10 diagnosis coding, which this article aims to do.

Previous research [15] has evaluated the model eXtreme Gradient Boosting (XGBoost) against the models Random Forest (RF) and Support Vector Machines (SVM) in the ability to predict sarcasm in natural text using punchline utterance and context. In the article, it was found that XGBoost achieved a higher F_1-score than RF and SVM when using only utterance as well as when using both utterance' and context. In the study, LIME and SHAP were used to give explanations for individual predictions. By using LIME and SHAP, the study could show that the models can explain that word importance is vital to correctly predict sarcasm in dialogues.

In another article [16], a user study was performed to evaluate the performance of an XAI system called HealthXAI. HealthXAI had the purpose of predicting cognitive decline from early symptoms. In the study, participants performed evaluations on the three factors of User Trust and Reliance (UTR), Explanation Satisfaction (ES) and Human-Machine Task Performance (HMTP).

Eight neurologists (clinicians) participated in the study who were well versed with technology and experts in cognitive decline. The explanations provided by HealthXAI were evaluated through a questionnaire using Likert scale answers. The study showed that HealthXAI with explanations performed better for all three factors than without explanations. Furthermore, the participants were very positive toward the explanations by HealthXAI for all three metrics.

In a related study [17], a proposed model aimed at explaining local multi-label classifications in NLP was compared to LIME and XCPlain. The models were evaluated through a user study and found that users could complete tasks faster with recommendations from the proposed model than with LIME. Additionally, Hamming score was used to evaluate the models, which is the fraction of correctly predicted labels out of all labels. On one dataset, LIME achieved 91%, the proposed model 90.75% and CXPlain 81.67%. On another dataset, LIME achieved 66.08%, the proposed model 65.23% and CXPlain 52.95%.

A previous study [13] compared an attention-based Recurrent Neural Network (RNN) to a basic RNN on which SHAP has been applied in the ability to give local and global explanations of Adverse Drug Events (ADE) in Swedish medical records. In the study, users assessed the explanations by the attention-based RNN and SHAP. Also, the Top-k Jaccard Index was used to assess the explanations by comparing the index of the models to those of medical experts. The medical experts in the study thought that SHAP gave more efficient explanations to show how features additively contribute to predictions. As such, SHAP was deemed most suitable for real-time scenarios where efficiency is important.

As apparent by the related research described above, there is previous research that has investigated the explainable performance of LIME and SHAP. However, there have been no comparisons of LIME and SHAP in their ability to explain ICD-10 classifications of Swedish gastrointestinal discharge summaries.

3 METHODOLOGY

3.1 Hypotheses

As stated, the aim of this study is to compare LIME and SHAP for the factors of user trust, explanation satisfaction and perceived usability. We hypothesise a difference between LIME and SHAP in terms of the three aforementioned factors for explaining ICD-10 classifications of Swedish gastrointestinal discharge summaries made by SweDeClin-BERT.

3.2 Selection of XAI Approaches

There is a multitude of XAI models that could be evaluated in explaining ICD-10 classifications. In a recent systematic review [8], 137 papers proposing XAI models were reviewed. However, not all the models can be considered for our study. For this study, the XAI models need to be local and post hoc. This means that they can explain individual classifications and be applied to existing prediction models respectively [18][19]. These two factors are necessary for an XAI model to be implemented in a CAC-tool for ICD-10 coding. The XAI must be able to explain individual classifications of ICD-10 codes and need to have the versatility of being applied to powerful classification models like BERT. This reduces the 137 models reviewed in [8] to 51 models. Further delimitation has been made by ranking the 51 remaining models by citations on Google Scholar. The model Gradient-weighted Class Activation Mapping (Grad-CAM) with 8,134 citations can be disregarded as it is intended for computer vision. This leaves LIME and SHAP are the most relevant models, with 8,430 and 6,432 citations, respectively, as of 2022-04-01. We use the number of citations to judge which models are most used and use this as a proxy for relevance.

3.3 Collection of Data for ICD-10 Classification

The data used in this study consist of Swedish gastrointestinal discharge summaries contained in the second version of the Stockholm EPR Gastro ICD-10

Corpus (ICD-10 Corpus) [1]. The ICD-10 Corpus is part of the research infrastructure Health Bank at DSV/Stockholm University. The Health Bank[2] contains Swedish patient records from over 2 million patients from Karolinska University Hospital, encompassing the years 2006 to 2014 [20]. All data in our study have been de-identified and hereafter called Stockholm EPR Gastro ICD-10 Pseudo Corpus or ICD-10 Pseudo Corpus for short.

The ICD-10 Pseudo Corpus consists of 6,014 gastro-intestinal discharge summaries. The dataset has a heavily imbalanced distribution of ICD-10 codes and this can be seen in Figure 1. To have a meaningful evaluation of LIME and SHAP, the predictions of ICD-10 codes being explained need to be made from a high-quality classifier. To mitigate the negative impact of the imbalanced data, a subset selection of discharge summaries has been made that have at least one of 18 selected ICD-10 codes out of the 263 original ones. While this approach might not be appropriate for an end-product application, it allows us to simulate a scenario where LIME and SHAP are applied to a model that has learned the underlying patterns of the data. This enables LIME and SHAP to also learn the underlying patterns of the data as they try to approximate the prediction function of the classification model. The subset selection of discharge summaries means that there are at least 100 discharge summaries for each of the 18 selected ICD-10 codes. The other discharge summaries, which do not have one of the 18 ICD-10 codes, have been removed from the subset. The subset consists of 3,636 samples. The distribution of ICD-10 codes in the subset and the ICD-10 codes are visible in Figure 2.

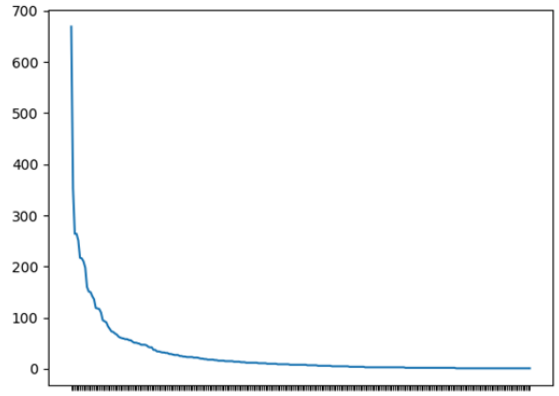

Figure 1. ICD-10 code distribution in the original dataset of ICD-10 Pseudo corpus. There are in total 263 unique ICD-10 codes on the X-axis.

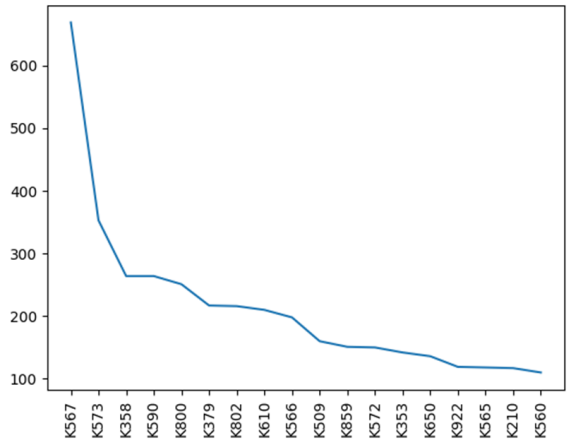

Figure 2. ICD-10 code distribution in the subset of 18 ICD-10 diagnosis codes.

As is typical in machine learning [21], the data has been split into a training set, validation set and test set. The training set consists of 2,617 samples (72%), the validation set of 655 samples (18%), and the test set of 364 samples (10%).

3.4 Implementation of SweDeClin-BERT

SweDeClin-BERT [11] has been used as the classification model in this study, a model based on the general Swedish KB-BERT [22] that has been further pre-trained on pseudonymised clinical text from the Health Bank. Pseudonymised means sensitive personal information has been identified in the text and replaced with surrogate values. SweDeClin-BERT is, therefore, a privacy preserving clinical language model for Swedish.

In our study SweDeClin-BERT has been fine-tuned for the downstream task of labelling discharge summaries with ICD-10 codes, using the aforementioned dataset of 3,636 discharge summaries. The fine-tuning has been done with the hyperparameters described in Table 1. To determine the number of epochs to train for, 5-fold cross-validation has been performed. The validation loss can be seen in Figure 3, resulting in the decision to fine-tune the model for nine epochs.

Hyperparameter Name	Value
Batch size	2
Learning rate	2e-5
Gradient accumulation steps	16
Number of warmup steps	155
Weight decay	0.01

Table 1. Hyperparameters for fine-tuning SweDeClin-BERT

[1] This research has been approved by the Regional Ethical Review Board in Stockholm under permission no. 2007/1625-31/5.

[2] Health Bank, http://dsv.su.se/healthbank

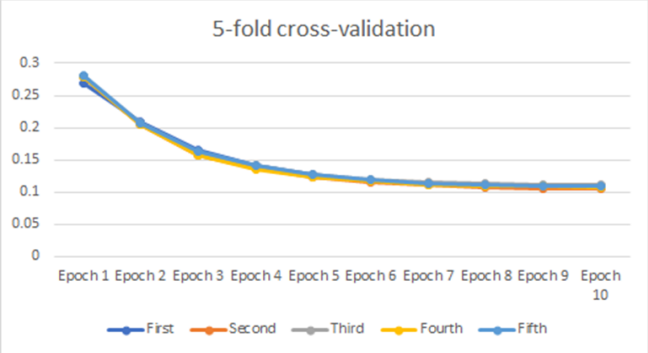

Figure 3. Validation loss during the five folds of the 5-fold cross validation

All hyperparameters except the number of epochs are based upon earlier studies [12]. Since the goal of this study is not to create an optimal classification model, no hyper-parameter optimisation has been performed. The code used to implement the model can be found in our Github repository[3].

3.5 Evaluation Results of SweDeClin-BERT

As SweDeClin-BERT returns prediction probabilities for each of the labels to a discharge summary, labelling has been considered true when having a prediction probability of 0.5 or higher. All evaluation results have been rounded to two decimals. The evaluation of SweDeClin-BERT on the test set has returned a mean accuracy of 0.97, as well as the mean results in Table 2. Micro and macro averaging captures different things. Micro averaging gives equal weight to every sample in a dataset, while macro gives equal weight to every class [23].

	Precision	Recall	F_1
Weighted	0.95	0.97	0.96
Micro averaged	0.97	0.97	0.97
Macro averaged	0.75	0.76	0.75

Table 2. Evaluation results of SweDeClin-BERT

3.6 Implementation of LIME and SHAP

In Figures 4 and 5, as well as Figure 6, one of the de-identified discharge summaries used in the questionnaire can be seen explained by LIME [24] and SHAP [25], respectively.

When LIME was implemented for this study, ten features were used and 100 samples. As such, ten is the greatest number of features for an explanation and 100 is the size of the neighbourhood of closest samples used to learn the linear model [26]. Ten features are the default value of the library, while 100 samples have been chosen due to computational limitations. On the left-hand side of Figure 4, LIME lists the prediction probabilities for the most probable ICD codes for the particular discharge summary. On the right-hand side, LIME lists the features that have the strongest influence for classifying the discharge summary with a certain ICD code. In Figure 5, the features can be seen highlighted in the discharge summary. The feature with the darkest colour has the highest impact. In Figure 6, the same discharge summary can be seen explained by SHAP. It has been classified with the ICD-code K859, with $f_{K859}(inputs)$ value of 1.96426. $f_{outputclass}(inputs)$ is the output from the model for the original output [27]. The base value for K859 is -1.03703, which means that it is the average prediction for that label [28]. Similarly to the visualisation by LIME, a darker colour indicates a more impactful feature. The blue features impact the classification negatively, while the red features impact it positively. In Figure 6, the ten most important features have been toggled to show their SHAP value. In the SHAP tool, more features can be toggled at will.

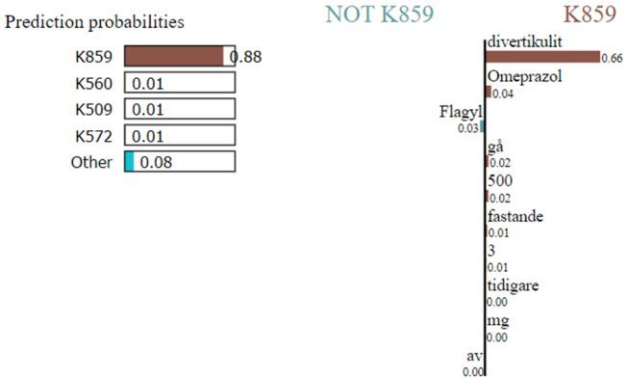

Figure 4. Example of LIME explanation in Swedish - prediction probabilities and features

3.7 Questionnaire Design

15 randomly selected discharge summaries from the test set have been included in the questionnaire, where the explanations for their predicted ICD-10 codes are explained by LIME and SHAP. This set of discharge summaries includes samples whose ICD-10 codes have been correctly predicted and ones that have been incorrectly predicted. This choice has been made to not give a misrepresentative view of the AI model's performance. Only the ten most

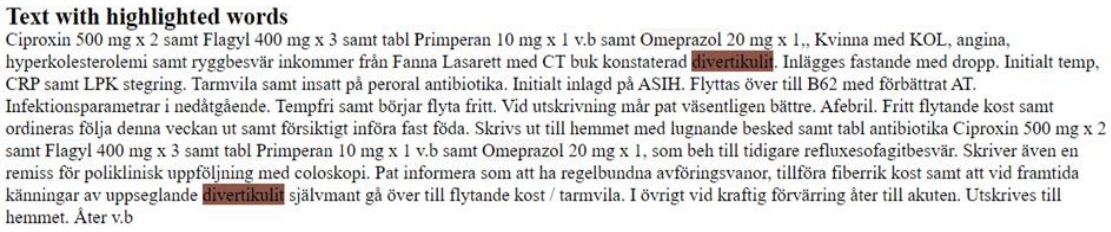

Figure 5. Example of LIME explanation in Swedish – features highlighted in text

[3] https://github.com/Alex01234/MastersThesis

Figure 6. Example of SHAP explanation in Swedish

impactful features for the most probable ICD-10 code from LIME and SHAP's explanations have been included in the questionnaire. The explanations by LIME and SHAP are evaluated by the respondents through the questionnaire in Google Forms.

As the explanations by LIME and SHAP are visually very different, the explanations have been harmonised in order to reduce potential design preference bias by the respondents. See examples of harmonised explanations in Figures 7 and 8, as well as Figures 9 and 10, for LIME and SHAP, respectively. In the harmonised explanations, the positively contributing features to an ICD-10 classification have a green colour. The negatively contributing features have a red colour. Again, the gastrointestinal discharge summaries are in Swedish. For reference, contrast the harmonised explanations with their original counterparts seen in section 3.6.

Prediction probability of **K859** is **0.88**, 10 most significant features:

NOT K859	K859	Weights
	divertikulit	0.66
	Omeprazol	0.04
Flagyl		0.03
	gå	0.02
	500	0.02
	fastande	0.01
	3	0.01
	tidigare	0.00
	mg	0.00
av		0.00

Figure 7. Harmonised explanation in Swedish by LIME - features and weights

Ciproxin 500 mg x 2 samt Flagyl 400 mg x 3 samt tabl Primperan 10 mg x 1 v.b samt Omeprazol 20 mg x 1,, Kvinna med KOL, angina, hyperkolesterolemi samt ryggbesvär inkommer från Fanna Lasarett med CT buk konstaterad divertikulit. Inlägges fastande med dropp. Initialt temp, CRP samt LPK stegring. Tarmvila samt insatt på peroral antibiotika. Initialt inlagd på ASIH. Flyttas över till B62 med förbättrat AT. Infektionsparametrar i nedåtgående. Tempfri samt börjar flyta fritt. Vid utskrivning mår pat väsentligen bättre. Afebril. Fritt flytande kost samt ordineras följa denna veckan ut samt försiktigt införa fast föda. Skrivs ut till hemmet med lugnande besked samt tabl antibiotika Ciproxin 500 mg x 2 samt Flagyl 400 mg x 3 samt tabl Primperan 10 mg x 1 v.b samt Omeprazol 20 mg x 1, som beh till tidigare refluxesofagitbesvär. Skriver även en remiss för poliklinisk uppföljning med coloskopi. Pat informera som att ha regelbundna avföringsvanor, tillföra fiberrik kost samt att vid framtida känningar av uppseglande divertikulit självmant gå över till flytande kost / tarmvila. I övrigt vid kraftig förvärring åter till akuten. Utskrives till hemmet. Åter v.b

Figure 8. Harmonised explanation in Swedish by LIME - features highlighted in text

Base value = -1.03703
f_{K859}(inputs) = 1.96426

NOT K859	K859	Weights
	buk konstaterad divertikulit.	2.746
	av uppseglande divertikulit självmant	1.424
refluxesofagitbesvär. Skriver även		-0.903
	Inlägges fastande med dropp.	0.796
en remiss för poliklinisk		-0.297
b samt Omeprazol 20		-0.257
fast föda. Skrivs		-0.224
ut till hemmet med		-0.218
x 2 samt Flagyl		-0.165
antibiotika Ciproxin 500 mg		-0.161

Figure 9. Harmonised explanation in Swedish by SHAP - features and weights

Ciproxin 500 mg x 2 samt Flagyl 400 mg x 3 samt tabl Primperan 10 mg x 1 v.b samt Omeprazol 20 mg x 1,, Kvinna med KOL, angina, hyperkolesterolemi samt ryggbesvär inkommer från Fanna Lasarett med CT buk konstaterad divertikulit. Inlägges fastande med dropp. Initialt temp, CRP samt LPK stegring. Tarmvila samt insatt på peroral antibiotika. Initialt inlagd på ASIH. Flyttas över till B62 med förbättrat AT. Infektionsparametrar i nedåtgående. Tempfri samt börjar flyta fritt. Vid utskrivning mår pat väsentligen bättre. Afebril. Fritt flytande kost samt ordineras följa denna veckan ut samt försiktigt införa fast föda. Skrivs ut till hemmet med lugnande besked samt tabl antibiotika Ciproxin 500 mg x 2 samt Flagyl 400 mg x 3 samt tabl Primperan 10 mg x 1 v.b samt Omeprazol 20 mg x 1, som beh till tidigare refluxesofagitbesvär. Skriver även en remiss för poliklinisk uppföljning med coloskopi. Pat informera som att ha regelbundna avföringsvanor, tillföra fiberrik kost samt att vid framtida känningar av uppseglande divertikulit självmant gå över till flytande kost / tarmvila. I övrigt vid kraftig förvärring åter till akuten. Utskrives till hemmet. Åter v.b

Figure 10. Harmonised explanation in Swedish by SHAP - features highlighted in text

All 30 explanations (15 discharge summaries explained by both LIME and SHAP) have three questions attached to them to evaluate the explanations on the factors of user trust, explanation satisfaction and perceived usability. The three questions are as follows:

- On a scale from 1 to 5, how trustworthy do you find the explanation of sample x to be?
- On a scale from 1 to 5, how satisfied are you with the explanation of sample x?
- On a scale from 1 to 5, how useful would you find the explanation of sample x to be, if used as a recommendation to classify the discharge summary?

4 RESULTS

Answers to the questionnaire have been collected from eight respondents, where seven are medical doctors/physicians, and one is a professional ICD-coder. All respondents have experience with ICD-coding. The collected data through the questionnaire has resulted in 120 data points for each of the factors of user trust, explanation

satisfaction and perceived usability for both models. The 120 data points come from the number of respondents multiplied by the number of discharge summaries in the questionnaire.

Paired t-tests have been performed to compare the score between LIME and SHAP for the aforementioned factors. Normally, some assumptions need to hold for a paired t-test [29]. These assumptions can, however, be disregarded when using Likert scale data [30].

4.1 Test for User Trust

A paired t-test has been conducted with the user trust for LIME and the user trust for SHAP, instantiated in the variables LIME_UT and SHAP_UT, respectively. The results can be seen in Tables 3 and 4. The *p*-value is 0.012, meaning that the difference of the means between LIME_UT and SHAP_UT is statistically different from zero at $\alpha = 0.05$ level of significance. The mean user trust for SHAP ($M = 2.99$, $SD = 1.553$) is higher than the mean user trust for LIME ($M = 2.47$, $SD = 1.478$), $t(119) = -2.544$, $p = 0.012$.

	Mean	N	Std. Deviation	Std. Error Mean
LIME_UT	2.47	120	1.478	.135
SHAP_UT	2.99	120	1.553	.142

Table 3. Paired samples statistics of LIME_UT and SHAP_UT

	LIME_UT - SHAP_UT
Mean	-.525
Std. Deviation	2.260
Std. Error Mean	.206
95% Confidence Interval of the Difference – Lower	-.934
95% Confidence Interval of the Difference – Upper	-.116
t	-2.544
df	119
Significance – One-Sided p	.006
Significance – Two-Sided p	.012

Table 4. Paired samples test of LIME_UT and SHAP_UT

4.2 Test for Explanation Satisfaction

A paired t-test has also been conducted with the explanation satisfaction for LIME and the explanation for SHAP, instantiated in the variables LIME_ES and SHAP_ES, respectively. The results can be seen in Tables 5 and 6. The *p*-value is 0.002, meaning that the difference of the means between the variables LIME_ES and SHAP_ES is statistically different from zero at $\alpha = 0.05$ level of significance. The mean explanation satisfaction for SHAP ($M = 3.04$, $SD = 1.434$) is higher than the mean explanation satisfaction for LIME ($M = 2.44$, $SD = 1.377$), $t(119) = -3.191$, $p = 0.002$.

	Mean	N	Std. Deviation	Std. Error Mean
LIME_ES	2.44	120	1.377	.126
SHAP_ES	3.04	120	1.434	.131

Table 5. Paired samples statistics of LIME_ES and SHAP_ES

	LIME_ES - SHAP_ES
Mean	-.600
Std. Deviation	2.060
Std. Error Mean	.188
95% Confidence Interval of the Difference – Lower	-.972
95% Confidence Interval of the Difference – Upper	-.228
t	-3.191
df	119
Significance – One-Sided p	<.001
Significance – Two-Sided p	.002

Table 6. Paired samples test of LIME_ES and SHAP_ES

4.3 Test for Perceived Usability

A paired t-test has been carried out with the perceived usability for LIME and the perceived usability for SHAP, instantiated in the variables LIME_PU and SHAP_PU, respectively. The results can be seen in Tables 7 and 8. The *p*-value is 0.005, meaning that the difference of the means between the variables LIME_PU and SHAP_PU is statistically different from zero at $\alpha = 0.05$ level of significance. The mean perceived usability for SHAP ($M = 2.99$, $SD = 1.569$) is higher than the mean perceived usability for LIME ($M = 2.39$, $SD = 1.485$), $t(119) = -2.855$, $p = 0.005$.

	Mean	N	Std. Deviation	Std. Error Mean
LIME_PU	2.39	120	1.485	.136
SHAP_PU	2.99	120	1.569	.143

Table 7. Paired samples statistics of LIME_PU and SHAP_PU

	LIME_PU - SHAP_PU
Mean	-.600
Std. Deviation	2.302
Std. Error Mean	.210
95% Confidence Interval of the Difference – Lower	-1.016
95% Confidence Interval of the Difference – Upper	-.184
t	-2.855
df	119
Significance – One-Sided p	.003
Significance – Two-Sided p	.005

Table 8. Paired samples test of LIME_PU and SHAP_PU

4.4 Analysis

SHAP has a higher mean value than LIME for all three factors investigated in our study. As evident from the results, SHAP has a mean value of around 3.0 for all factors, while LIME has a mean value of ca 2.4 for all factors. It could therefore be worthwhile to further study the explainable performance of SHAP. This could be particularly interesting since the explanations by SHAP have been harmonised with the explanations by LIME in this article. Further studies could evaluate SHAP in its original format, where the full capability of the model can be utilised.

5 DISCUSSION

This article has aimed to evaluate relevant XAI models that could be incorporated into a CAC-tool for ICD-10 coding. The most relevant XAI models for this purpose have been judged to be LIME and SHAP. While there exist previous studies that have evaluated LIME and SHAP on NLP tasks, there have been few studies evaluating them in a Swedish context. Previous research has given indications of the promising potential of LIME and SHAP. One example is LIME and SHAP's ability to show that word importance is crucial to predicting sarcasm in dialogues. Another one is that medical experts think that SHAP gives efficient explanations of how features additively contribute to explanations when explaining ADE in Swedish medical records. However, this paper is the first one to compare the ability of LIME and SHAP to explain the assignment of ICD-10 diagnosis codes to Swedish discharge summaries.

The main limitation of our study is its generalisability. A non-probabilistic approach has been applied to recruit respondents to generate an exploratory sample. This means that the opinions of the respondents may not be representative of the whole research population. Furthermore, the evaluations of LIME and SHAP are heavily dependent on the performance of the underlying model on which it is applied, in this case, SweDeClin-BERT. The evaluations are also dependent on the data used, in this case, Swedish gastrointestinal discharge summaries. This is since LIME and SHAP try to approximate the prediction function of the model to which they are applied.

If a CAC-tool for ICD-10 coding is to be constructed in the future, there are many things that will have to be optimised in comparison to what has been done in this article. For example, a balanced dataset will have to be used, contrary to the dataset used in this article. Additionally, the hyperparameters of the classification model will have to be optimised during training. Once satisfactory predictive performance on all ICD-10 codes has been established, the results of this study can be used as decision support on which XAI model to incorporate into the CAC-tool. Naturally, the limitations of this study have to be kept in mind when making such a decision.

Future research is recommended to conduct a similar survey with a larger sample, which could have greater generalisability for the whole research population. Future research could also be done that applies LIME and SHAP on a classification model that has been optimised, using the considerations in the previous paragraph. The data used could also be extended to not only include gastrointestinal discharge summaries. Other kinds of discharge summaries, as well as other medical records than only discharge summaries, could be used to train the underlying classification model on which an XAI model is applied. If this increases the predictive capability of the underlying classification model, it will likely improve the explanations by the applied XAI model. Furthermore, a qualitative study is recommended with the original SHAP tool, where visualisations are unaltered (as in Figure 6 rather than Figure 9 and 10). Then all the aspects of SHAP's visualisations can be evaluated to gain knowledge of which aspects of the explanations are valuable for a future CAC-tool. This could be especially interesting, as medical experts in previous research have indicated that SHAP gives efficient explanations of how features additively contribute to predictions. Such a qualitative study does not have to be limited to SHAP only, as it could be worthwhile to investigate the full capability of LIME as well.

In this article, explanations are only given for the most likely predicted ICD-10 code. However, a CAC-tool might include a longer list of suggested codes. A qualitative study where recommendations are given for multiple ICD-10 codes could also uncover interesting findings.

6 CONCLUSIONS

This article has examined and compared the explainable performance of LIME and SHAP in their ability to explain ICD-10 classifications of Swedish gastrointestinal discharge summaries. The classification model SweDeClin-BERT has been fine-tuned for the task of labelling discharge summaries with ICD-10 codes. LIME and SHAP have then been applied to SweDeClin-BERT to generate explanations for SweDeClin-BERTs classifications. 15 discharge summaries have been randomly chosen from the test set of data and visualised in a questionnaire. In the questionnaire, the ten most impactful features for the most probable ICD-10 code as deemed by LIME and SHAP have been visualised. The visualisations have been harmonised to reduce design preference bias. Eight answers have been collected from respondents experienced in ICD-10 coding, who have evaluated the explanations by LIME and SHAP by the factors of user trust, explanation satisfaction and perceived usability. The results of paired t-tests show that there is a statistically significant difference between LIME and SHAP for the mean value of all factors. SHAP has a higher mean value than LIME for all three factors.

7 REFERENCES

[1] World Health Organization. 2022. International Statistical Classification of Diseases and Related Health Problems (ICD). https://www.who.int/standards/classifications/classification-of-diseases (Accessed 2022-05-27)

[2] World Health Organization. 2022. Importance of ICD. https://www.who.int/standards/classifications/frequently-asked-questions/importance-of-icd (Accessed 2022-07-02)

[3] Jacobsson A., Serde, L. 2013. Kodningskvalitet i patientregistret (In Swedish). https://www.socialstyrelsen.se/globalassets/ sharepoint-dokument/artikelkatalog/statistik/2013-3-10.pdf (Accessed 2022-05-27)

[4] Norwegian Centre for E-health Research. 2022. ClinCode - Computer-Assisted Clinical ICD-10 Coding for improving efficiency and quality in healthcare. https://ehealthresearch.no/en/projects/clincode-computer-assisted-clinical-icd-10-coding-for-improving-efficiency-and-quality-in-healthcare (Accessed 2022-07-02)

[5] Pu, Y., Apel, DB., Liu, V. Mitri, H. 'Machine learning methods for rockburst prediction-state-of-the-art review', in *International Journal of Mining Science and Technology*. Vol. 29, Issue. 4, pp. 565 – 570, 2019.

[6] Young, T., Hazarika, D., Poria, S., Cambria, E. 'Recent trends in deep learning based natural language processing', in *ieee Computational intelligenCe magazine*. Vol. 13, Issue. 3, pp. 55-75, 2018.

[7] Asan, O., Bayrak, A. E., Choudhury, A. 'Artificial intelligence and human trust in healthcare: focus on clinicians', in *Journal of medical Internet research*. Vol. 22, Issue. 6, e15154, 2020.

[8] Islam, MR., Ahmed, MU., Barua, S., Begum, S. 'A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks', in *Appl. Sci*. Vol. 12, Issue. 3, p. 1353, 2022.

[9] Ribeiro, MT., Singh, S., Guestrin, C. '"Why should I trust you?" Explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144, 2016.

[10] Lundberg, SM., Lee, SI. 'A unified approach to interpreting model predictions', in *Advances in neural information processing systems, 30*, 2017.

[11] Vakili, T., Lamproudis, A., Henriksson, A., Dalianis, H. 'Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data', accepted to *LREC 2022*, 2022.

[12] Devlin, J., Chang, MW., Lee, K., Toutanova, K. 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*, 2018.

[13] Rebane, J., Samsten, I., Pantelidis, P., Papapetrou, P. 'Assessing the Clinical Validity of Attention-based and SHAP Temporal Explanations for Adverse Drug Event Predictions', in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 235-240, 2021.

[14] Remmer, S., Lamproudis, A., Dalianis, H. 'Multi-label Diagnosis Classification of Swedish Discharge Summaries–ICD-10 Code Assignment Using KB-BERT', in *RANLP 2021: Recent Advances in Natural Language Processing, 1-3 Sept 2021, Varna, Bulgaria*, pp. 1158-1166, 2021.

[15] Kumar, A., Dikshit, S., Albuquerque, VHC. 'Explainable Artificial Intelligence for Sarcasm Detection in Dialogues', in *Wireless Communications and Mobile Computing, 2021*, 2021.

[16] Khodabandehloo, E., Riboni, D., Alimohammadi, A. 'HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline', in *Future Generation Computer Systems*, Vol. 116, pp. 168-189, 2021.

[17] Singla, K., Biswas, S. 'Machine learning explanability method for the multi-label classification model', in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pp. 337-340, 2021.

[18] Vilone, G., Longo L. 'Explainable artificial intelligence: a systematic review', *arXiv preprint arXiv:2006.00093*, 2020.

[19] Vilone, G., Longo L. 'Classification of explainable artificial intelligence methods through their output formats', in *Machine Learning and Knowledge Extraction*, Vol. 3, Issue. 3, pp. 615-661, 2021.

[20] Dalianis, H., Henriksson, A., Kvist, M., Velupillai, S., Weegar, R. 'HEALTH BANK-A Workbench for Data Science Applications in Healthcare', in *CAiSE Industry Track.* Vol. 1381, pp. 1-18, 2015.

[21] James, Gareth. Witten, Daniela. Hastie, Trevor. Tibshirani, Robert. *An Introduction to Statistical Learning*. Springer, New York, p. 198, 2013.

[22] Malmsten, M., Börjeson, L., Haffenden, C. 'Playing with Words at the National Library of Sweden--Making a Swedish BERT', *arXiv preprint arXiv:2007.01658*, 2020.

[23] Yang, Y. 'An evaluation of statistical approaches to text categorization', in *Information retrieval*. Vol. 1, Issue. 1, pp.69-90, 1999.

[24] Ribeiro, MT. 2021. marcotcr/lime. https://github.com/marcotcr/lime (Accessed 2022-04-21)

[25] Lundberg, SM. 2022. SHAP. https://github.com/slundberg/shap (Accessed 2022-03-30)

[26] Ribeiro, MT. 2016. lime package. https://lime-ml.readthedocs.io/en/latest/lime.html (Accessed 2022-04-22)

[27] Lundberg, S. 2018. Using custom functions and tokenizers. https://shap.readthedocs.io/en/latest/example_notebooks/text_examples/sentiment_analysis/Using%20custom%20functions%20and%20tokenizers.html (Accessed 2022-05-28)

[28] Molnar, C. SHAP (SHapley Additive exPlanations). In *Interpretable Machine Learning*, 2022. https://christophm.github.io/interpretable-ml-book/shap.html (Accessed 2022-03-30)

[29] Newcastle University. n.d. Paired Samples T Test (Dependent Samples T test). https://services.ncl.ac.uk/itservice/research/dataanalysis/simpletests/ttests/pairedsamplesttestdependentsamplesttest/ (Accessed 2022-04-17)

[30] Norman, G. 'Likert scales, levels of measurement and the "laws" of statistics', in *Advances in health sciences education*, Vol. 15, Issue. 5, pp. 625-632, 2010.

8 ACKNOWLEDGEMENT

This study has been partially funded by the Norwegian Research Council through the ClinCode project, number 318098.

We would also like to thank our eight respondents for their work.