Identifying a Trial Population for Clinical Studies on Diabetes Drug Testing with Neural Networks

LÖHR Tim, B.Sc.
Friedrich Alexander University
tim.loehr@fau.de

***Abstract*—This project aims to model an end-to-end workflow of implementing Artificial Intelligence (AI) for the clinical environment. A possible use case such as the selection process of patients for a novel treatment or drug will be conducted by estimating the hospitalization time with a Neural Network. The diabetes readmission dataset from the University of California, Irvine (UCI) Diabetes is used for this project. The trial population is selected by predicting the expected days for a person being hospitalized. An arbitrary boundary is set for chosing whether or not a patient shall be included into the trial. If so, a clear explanation of how the prediction is calculated and additional possible risk factors will be given in order to make the workflow explainable. This project shows that given a proper explanatory approach, AI can be a useful tool for the modern clinical environment. The workflow finally reveals that AI can be a beneficial support tool for doctors, for example by effectively choose possibly suitable patients in the patient selection process.**

***Index Terms*—Machine Learning in the Industry 4.0, Clinical EDA, Data Analysis, AI in Medicine, Neural Networks, Diabetes, FAU, Department of Computer Science**

1. Introduction

Due to the upcoming new law enforced by the German Health Minister *Jens Spahn*, electronic patient files will be nationwide standardized in Germany [1] on the first of January 2021. This opportunity can be used to increase the impact of AI on the health system to improve the health of Germany's population and facilitate doctor's work. Electronic Health Records (EHR) offer a variety of different application fields [2] [3], such as:

- Discover novel disease treatments.
- Improve patients diagnosis.
- Improve personalized healthcare.

As of today in 2020, there are seven million cases of diabetes in Germany, stated from the Robert Koch Institut (RKI) [4]. Even for well known deceases like diabetes, there is continuous research and novel drugs and treatments are invented on a frequent basis. Still, not every person is suitable for obtaining a novel treatment.

The project's background is an artificial scenario. The assumption is that the data is derived from the database collected through the new central storage system of the patient files due to the new German law. EHR are under high data privacy conditions. They can only be used, if the data is completely anonymized and patients can not be traced back. So it is assumed that those issues have already been takes care of and therefore it is pretended that all limitations and regularities have been successfully acquired.

After preparing the EHR for this privacy requirements, the data can be utilized to gain insights and evaluations out of it. Using AI in real-world clinics is a highly classified realization, because wrong decisions can cause a dangerous outcome towards patients. For this reason, implementing *AI in Medicine* requires a very detailed explanatory analysis of the predicted outcome. Keeping the background in mind, this project aims to implement an end-to-end workflow of how AI is used in the real clinical environment.

The first step is to figure out which patients are most likely suitable for receiving novel diabetes treatments. This can be achieved by a patient selection process, as stated by Dr. Toddenroth et. al. from the FAU [5] and a more recent paper from Szu-Yeu Hu et. al. [6]. Then the uncertainty estimation of the prediction is computed with

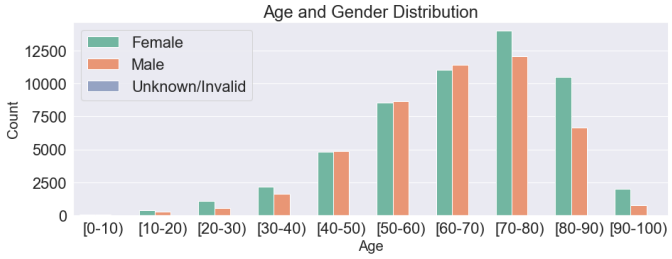

Fig. 1. The bivariate diagram shows the age disparity among all patients within the dataset, further split into the two gender bins.

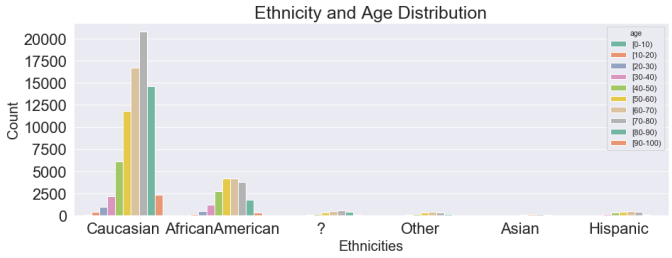

Fig. 2. The bivariate diagram shows the ethnicity disparity among all patients within the dataset, further split into single age bins.

the Aequitas tool from the University of Chicago. This estimation is performed to explain the predictions with Explainable Artificial Intelligence (XAI) methods. This process is supposed to enable an understanding of how certain estimations of Neural Networks are with respect to different dataset balance problems, e.g. demography. The research question for this project can therefore be concluded as:

Can Machine Learning be safely applied in the real clinical environment if it provides enough explainability for its predictions?

2. Methods

The dataset is originated from the UCI Machine Learning Repository [7] collected from 1999 to 2008 with over 100000 entries to train on. Various features such as demographics, diabetes conditions or medications are provided. In total there are 55 features from which 36 are included into the modeling, because the other 19 features were useless due to missing features and NaN values. An important detail is the distribution of ethnicity, gender and age. This disparity is used in section results for evaluating the uncertainty of the model. Figure 1 and 2 show that the median age of diabetes patients is 70 to 80 years. As can be seen from Figure 1, the gender can be ignored, because it is almost equally distributed among all patients. The ethnicity disparity reveals the biggest concern, because among different ethnicities there is a huge distribution gap. Asian and Hispanic people are very underrepresented. African American people in comparison to Caucasian people as well. This plays an important role in the later predictive analysis and explanatory approach. Keeping these distributions in mind, the methods used for this project will be conducted in the following respective order:

- Model a Neural Network.
- Explain the model with uncertainty estimation and metrics (**why** are the model's predictions good).
- Explain the predictions with SHAP and LIME (**how** is the prediction computed).

SHapley Additive exPlanations (SHAP) is a python package for approximating the Shapely values. Those values can be used for various explanatory approaches, but for this project it is used for plots like in Figure 7. LIME is a python package used for the LIME XAI approach. [1]

Preprocessing is most of the time the first step before modeling a Neural Network. Various unnecessary features were needed to be eliminated and categorical features were transformed into a one-hot encoded representation in order to funnel them into the Network properly. A supervised Neural Network always has a response variable (label), in order to make a proper prediction by calculation the loss and doing backpropagation. The *time in hospitalization* feature is used as the label, because it reveals the severity of each patient's diabetes decease. Usually, the longer people need to be hospitalized after receiving treatments, the worse the condition of this person is. Predicting a low hospitalization time with a novel treatment can indicate compatibility with the patient and therefore he or she is a good match to be selected.

Neural Networks are able to achieve far better results than standard Machine Learning approaches like the Logistic Regression. This can be proven due to the universal approximation theorem. Neural Networks are less explainable than most Machine Learning techniques, because they are mostly seen as a *blackbox*. The Neural Network for this project is modeled with a Keras sequential function, consisting of Dense, Dropout, Dense Variational and a Distribution Lambda layers to estimate the hospitalization time with a certain probability. The Distribution Lambda layer from Tensorflow incorporates an ensemble of models which report the variability of the prediction. It can be viewed as taking the mean of various regressions and measure the qualitative difference between all outputs. Usually in clinical data those layers can be implemented to eliminate the risk of:

- Aleatoric Uncertainty: statisical uncertainty (*known unknowns*).
- Epistemic Uncertainty systematic uncertainty (*unknown unknowns*).

Aleatoric uncertainty will be evaluated with respect to the ethnicity distribution of the data. The neural network's prediction will be measured based on different evaluation metrics:

[1]More detailed explanations of these two approaches can be further read on the official documentations.
SHAP: https://shap.readthedocs.io/en/latest/
LIME: https://lime-ml.readthedocs.io/en/latest/

- Precision, Recall and F1 Score.
- Brier Score.
- AUC Score.

The loss function for the backpropagation is the negativ log likelihood function [8]:

$$\text{Loss} = -log(y) \text{ for each prediction } y$$

The loss generally increases when the regression is unclear about its prediction and shrinks with increasing certainty. At this point, the model is able to predict how long a person is going to stay in the hospital with respect to the demographics. The loss has a direct influence on the performance of the metrics. Generally speaking, if the loss is zero, all the metrics are supposed to have a perfect score.

Conclusively to guarantee full transparency of the patient selection process, Shapely values and LIME try to explain which features led to which output. In other words, which features have a positive and a negative impact on the hospitalization time. The key difference between these two XAI methods is that Shapely values want to explain *how* one feature contributes to the overall prediction, while leaving out other combinations of features. LIME wants to explain what the most important features are for the prediction in general.

3. Results

After having the workflow in mind, the first step is to train the neural network and evaluate its results. The model trained for 50 epochs and reached a loss of 2.97. Figure 6 shows the course of events for different evaluation metrics. For example, the *Brier Score* itself is best when it is as low as possible. The other scores are best, when they are as close to 100 percent as possible. The X-Axis represents the predicted time of hospitalizations in the unit of measurement days. All metrics have a optimal score for the boundary at day four, as can be seen in Figure 6. After selecting the most promising discrimination boundary at day four, the uncertainty estimation with respect to ethnicity can be computed.

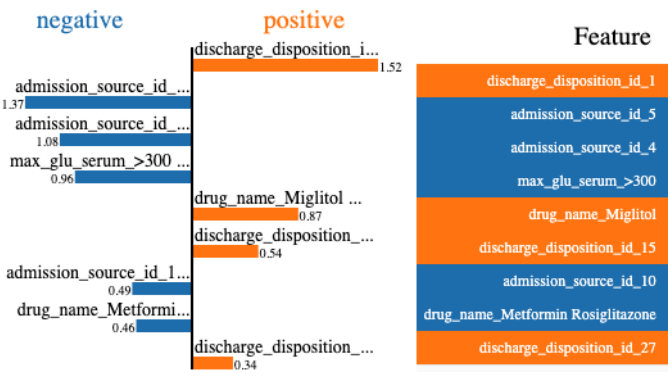

Fig. 3. Important features computed by LIME. The true label for this was seven days and model predicted a value of 6.42 days.

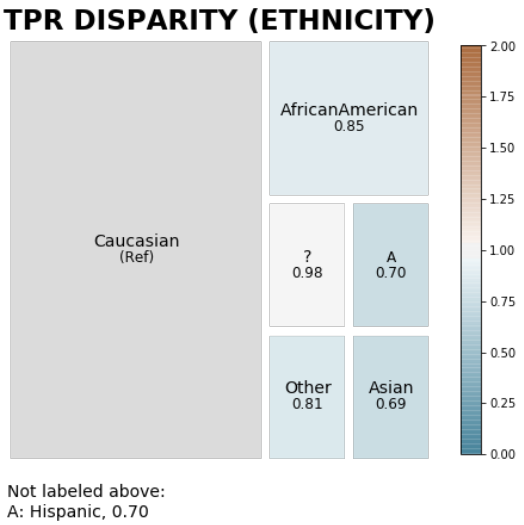

Fig. 4. True positive rate among ethnicities with Caucasiens as reference group with the discrimination boundary at day four.

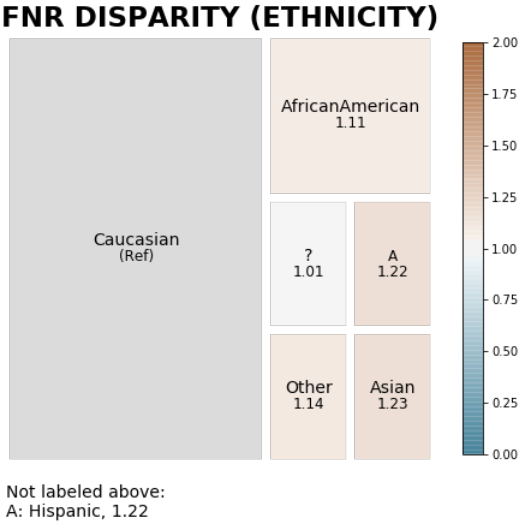

Fig. 5. False negative rate among ethnicities with Caucasiens as reference group with the discrimination boundary at day four.

Figures 4 and 5 highlight that the true positive rate for Caucasian people is higher compared to the other ethnicities. The false-negative rate of other ethnicities is higher compared to the reference Caucasian people.

The next step is the investigation of how each prediction is computed and what the most important features for those predictions are. SHAP and LIME are used for this purpose, as mentioned in the section methods.

The second interpretational approach are the Shapely values, which can be seen in Figure 7. This Figure lists 20 of the features with both their positive and negative impact on the prediction of the hospitalization time. Negative values on the X-Axis have an decreasing impact on the hospitalization time and positive values on the X-Axis will increase the time of hospitalizations.

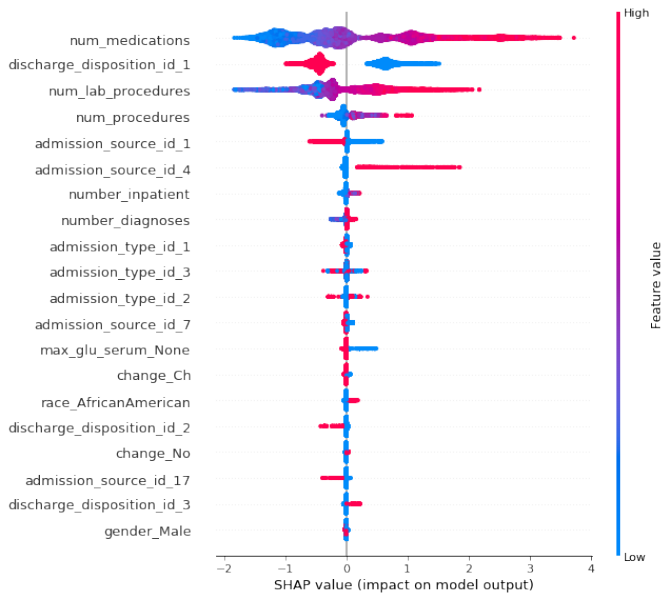

Fig. 7. False negative rate among ethnicities with Caucasiens as reference group with the discrimination boundary at day four.

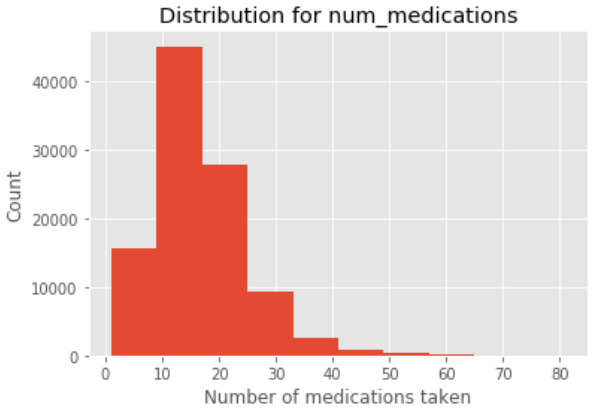

Fig. 8. Normal right skewed distribution of the feature *num_medications*.

4. Discussion

Training the model alone can not qualify the selection of a patient into the trial population. There needs to be a specific boundary indicating after which time of hospitalization the diabetes is so serious, that this person can be taken into consideration for the trial. If this boundary is not set, the selection process can not be accomplished. Figure 6 not only shows the course of events, but it also highlights that different boundaries for the X-Axis produce diverse scores. The optima is between the third and the firth day.

The uncertainty estimation can be evaluated for different demographics against each other and with different boundaries set on Figure 6. Figures 4 and 5 indicate, that if a patient is being selected and he or she is Caucasian, the probability of a positive or negative choice is more accurate compared to other ethnicities. Looking back at Figure 2, this important prediction disparity can be explained due to the huge difference in training data for all ethnicities.

Shapely values and LIME need to be interpreted in a different way than the metrics. Figure 3 illustrates the single components which led to this prediction of 6.42 days. The meaning of these categorical features are provided by the following bullet points:

- discharge_disposition_id_1: Discharged to home.
- admission_source_id_5: Transfer from a Skilled Nursing Facility (SNF).
- admission_source_id_4: Transfer from a hospital.
- admission_source_id_10: Transfer from critial access hospital.

This specific patient is marked with the features *admission_type_id_1* and discharge_disposition_id_6. This means that **not** having the features admission_source_id_5 and 6 has actually an **positive** (fewer days in hospital) impact on the hospitalization time, whereas when a patient has the *discharge_disposition_id_1*, this indicates a longer stay in the hospital. Also e.g. if the patient is taking the medication

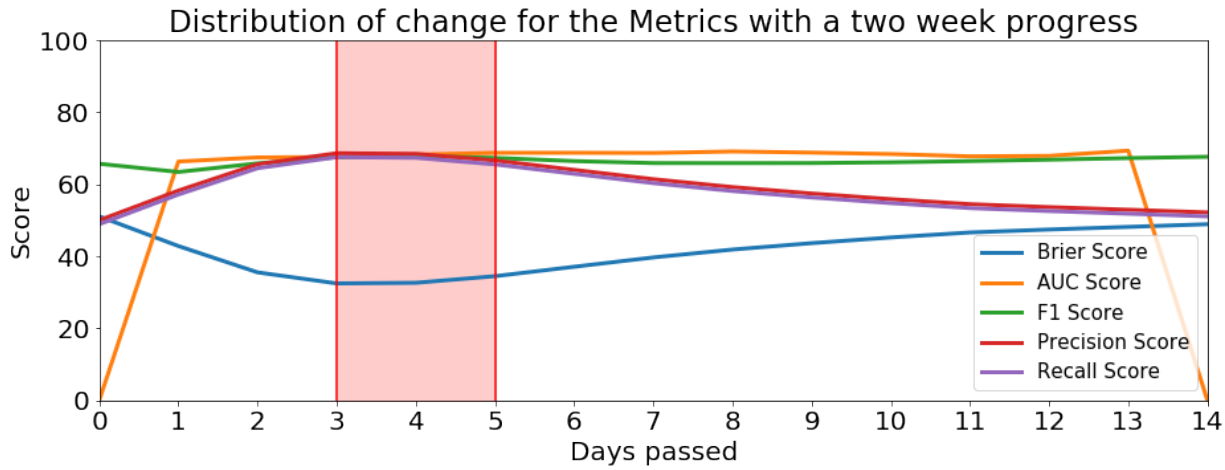

Fig. 6. The graphic shows the temporal change of the metrics with respect to a different time constraints on the boundary. The best values for the metrics can be found between the third and fifth day. Setting the boundary somewhere in the red highlighted area produces the best and most certain estimation.

Migitol, he or she generally will stay up to one day (0.87) longer in hospital than people who do not take this drug. Explanations like this can be conducted for all features to further investigate new insights of which effects medications and treatments can have.

The distribution of *num_medications* can be seen in Figure 8. This Figure shows a right-skewed normal distribution. Based on this distribution, it can be seen that a normally distributed numerical feature can have both a negative and a positive impact on the prediction of the most important features computed by SHAP, as shown in Figure 7. The more medications a person takes, the worse the condition and the longer the hospital stay for this person is, and vice versa. Whereas the feature *admission_source_id_4*, if marked positive for a person, has always a high positive impact (higher hospitalization time) for the patient.

After taking all evaluations and explanatory steps into consideration: Can Neural Networks safely be applied in the Industry 4.0 and especially hospitals? Keeping the results in mind, it appears that AI can have many good traits for applying it in hospitals.

Pros:

- Support for doctor's decision
- Patient selection saves a lot of time
- New insights of treatment of medication effects into the patients medical condition
- Possible new insights

Cons:

- Unstructured data
- Uncertainty if it can be applied among different EHR and not only diabetes data
- The minimum necessary needed data amount for highly explainable predictions is unclear

These techniques can be a good decision-support technology for doctors. If there are for example 2000 patients stored in the hospital's database, this pre-selection process can shrink the number of possible suitable people already down in advance. Doctors will make the final decision based on their experience, but they can end up saving a lot of time by looking at already e.g. 20 pre-selected patients instead of investigating 2000 patient files. So even if the prediction is not going to be selected, it has no dangerous impact for patients in general. Doctors should be taught to develop more trust into a well trained and good performing AI system, because it can be beneficial tool.

5. Conclusion

The conclusion of this entire workflow leads to three final questions to be answered.

Can it be applied without additional doctor's approval?

At this day and age not. What if only one person dies because of a false decision of the Neural Network's prediction? This is a heavily discussed topic and the law and regulatory clinical instances like the FDA in the United States or the Arzneimittelbehörde in Germany comes into force when implementing such technologies in the clinical environment.

Can it be applied as a support technology?

If a doctor validates the results, it can be used to support the clinical trials and save up a lot of time and identify new insights.

Does a combination of metrics improve the overall explainability and how many metrics are enough?

Using at first the F1-Score for evaluating the model's performance and then computing the uncertainty estimation with respect to critical features has an overall improvement. Furthermore it can be seen that many different approaches reveal different insights of the prediction. A good combination of metrics can enhance the overall performance.

6. Future Work

For future research further improvements can also be made, such as:

- Try different XAI methods or explain more features with different ways.
- Try different Neural Networks or classifiers with GPU for higher accuracies.
- Model also the Epistemic Uncertainty with different Neural Network Layers: tfp.layers.DenseVariational for a more robust Network.

References

[1] *Spahn: „Nur wenn wir die Chancen der Digitalisierung nutzen, koennen wir die Patientenversorgung besser machen“, author =.*

[2] M. E. H. H. A. G. Shahabeddin Abhari, Sharareh R. Niakan Kalhori, "Artificial intelligence applications in type 2 diabetes mellitus care: Focus on machine learning methods," Report, 2019.

[3] M. P. V. K. R. Amisha, Paras Malik, "Overview of artificial intelligence in medicine," Report, 2020.

[4] R. K. Institute, *Diabetes Surveillance*, 2020. [Online]. Available: https://www.rki.de/DE/Content/Gesundheitsmonitoring/Studien/Diabetes_Surveillance/Diabetesbericht.pdf?__blob=publicationFile

[5] R. F. A. S. C. N. M. S. R. C. H.-U. P. . D. T. Felix Köpcke, Dorota Lubgan, "Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data," Report, 2013.

[6] A. W. F. D. M. J. H. N. A. C.-D. B. K. R. B. J. A. T. C. L. Szu-Yeu Hu, Enrico Santus, "Can machine learning improve patient selection for cardiac resynchronization therapy?" Report, 2019.

[7] U. M. L. Repository, *Diabetes 130 US hospitals for years 1999-2008*, 1999-2008. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

[8] *Negative Log Likelihood Ratio Loss for Deep Neural Network Classification“, author =.*