FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

MACHINE LEARNING
& DATA ANALYTICS

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

**LÖHR Tim**
Friedrich Alexander University
*tim.loehr@fau.de*

***Abstract*—This project models an end-to-end workflow of implementing AI for the clinical environment. A possible use-case such as the selection of patients for a novel treatment or drug will be conducted, by estimating the hospitalization time with a Tensorflow regression Neural Network. Using a synthetic dataset from the UCI Diabetes readmission dataset, the expected days for a person being hospitalized after certain conditions or treatments will be predicted. This result is used to decide whether a patient is applicable to be included in the clinical trial. If so, there needs to be a clear explanation of the prediction and possible risk factors. This project shows the importance of splitting the data appropriately without data leakage and evaluating the results to make it transparent for the official use case, e.g. being accepted by the Arzneimittelbehörde or FDA as a decision support tool for hospitals or doctors.**

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

## 1. Introduction

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

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

As of today in 2020, there are 7 million cases of diabetes in Germany, stated from the Robert Koch Institut (RKI) [3]. Compared to Germany's population size of 83 million people, 7 million diabetes cases is a credible amount. Even for well known deceases like diabetes, there is continuous research and novel drugs and treatments are invented frequently. Still, not every person is suitable for obtaining a novel treatment.

The project's background an artificial scenario, in which I pretend to have received the data from the database collected through the new central storage system of the patient files due to the new law. This data is preserved with all privacy rights and anonymized to make use of it. This data can now be utilized to gain evaluations and therefore benefit AI research in the medical environment. Using AI in Medicine is a highly classified realization because wrong decisions can cause a dangerous outcome towards people. For this reason, implementing AI in Medicine requires a very detailed explanatory analysis for the predicted outcome. Keeping the background in mind, this project now aims to implement an end-to-end workflow of how to make use of AI 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 via a patient selection process, as stated by Dr. Toddenroth et. al. from the FAU [4] and a more recent paper from Szu-Yeu Hu et. al. [5]. Then the uncertainty estimation of the prediction is computed with the Aequitas tool from the University of Chicago. This is performed to explain with Explainable Artificial Intelligence (XAI) methods of how the prediction was determined. That process is supposed to enable an understanding of how certain and where AI can be used, e.g. as decision support.

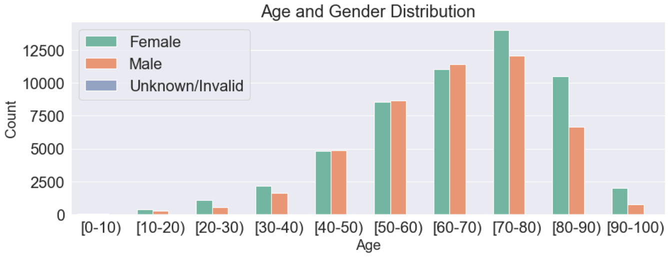

Fig. 1. Age and Gender Disparity within the dataset

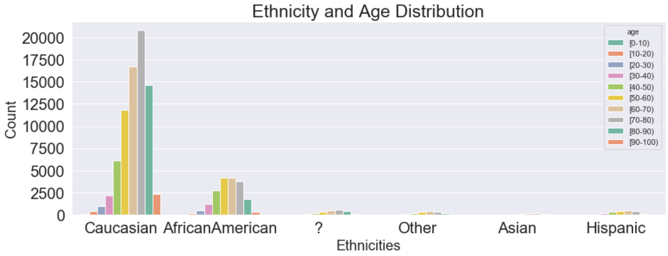

Fig. 2. Ethnicity Disparity within the dataset

The research question for this project can therefore be concluded as:

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

## 2. Methods

The dataset is originated from the UCI Machine Learning Repository [6] 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 for the modeling. An important detail is the distribution of ethnicity, gender and age. This disparity in distribution is used in section results for evaluating the uncertainty of the model.

Figure 1 and 2 show that the median age is around 70 to 80 years for diabetes patients, whereas gender can be mostly ignored, because it is almost equally distributed. The ethnicity disparity reveals a bigger concern, because among different ethnicities there is a huge distribution gap. Asians and Hispanics are very underrepresented and 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, lets with explaining the methods used for this project in respective order:

- Model a Neural Network
- Explain the model (**why**) with Uncertainty Estimation and Metrics
- Explain the predictions (**how**) with SHAP and LIME

Preprocessing is the first step for modeling a Neural Network. Various unnecessary features needed to be eliminated and categorical features needed to be transformed into a one-hot encoded representation in order to funnel them into the Network properly. The model needs a proper response variable, also called a label, in order to make a proper prediction for the patient selection. I decided on the *time in hospitalization* feature, 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 the person is. So predicting a low hospitalization time with a novel treatment can indicate compatibility with the patient and so he or she is a good match to be selected.

Neural Networks able to achieve far better results than standard Machine Learning approaches, due to the universal approximation theorem, yet they are less explainable. The Neural Network is modeled with a Keras sequential Neural Network, 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. Those layers are implemented to eliminate the risk of:

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

which usually occurs in clinical data. The aleatoric uncertainty will be evaluated with respect to the ethnicity distribution of the data. The epistemic uncertainty would break the limits of this project.

The neural network's prediction will be measured based on different metrics:

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

The loss function I used was the negativ log likelihood function:

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

The loss generally increases when the regression is unclear and shrinks with more certainty. At this point, the model is now able to predict how long a person is going to stay in the hospital for a specific condition and concerning that person's demographics. The loss has a direct influence on the performance of the metrics. Generally speaking, if the loss is 0, 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 those two XAI methods is that SHAPely wants to explain *how* the prediction

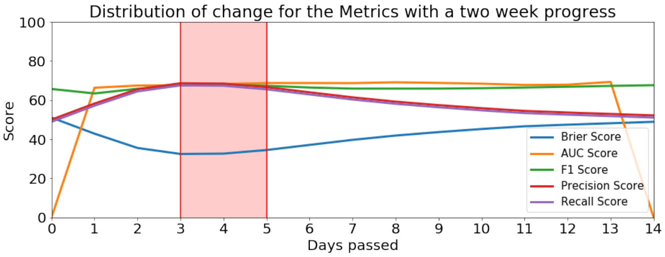

Fig. 3. The graphic shows the temporal change of the metrics with respect to a different time constraint on the boundary. The best values for the metrics can be found within the third and fifth day. Setting the boundary somewhere in the red highlighted area produces the best and most certain estimation.

was computed, whereas LIME wants to explain what the most important features were for the prediction.

## 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. Nevertheless, this alone can not qualify for selecting a patient. 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, otherwise the selection process can not be accomplished. Figure 3 shows the course of events for different evaluation metrics. For example the *Brier Score* is best when it is as low as possible, whereas the other scores are best when they are as close to 100 as possible. Taking this into consideration, if the boundary is set to four on the x-axis, which means four days of hospitalization, all metrics have the globally optimal score. The predictions of the model are closest to the true labels when the boundary is set to four. After carefully selecting the most promising discrimination boundary at four days for generating a label, the uncertainty estimation with respect to ethnicity can be computed.

Taking a close look at Figures 4 and 5, it can be seen that the true positive rate for Caucasian people is higher compared to the other ethnicities. Also, the false-negative rate of other ethnicities is higher compared to the reference Caucasian people. Both 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.

The next step is the investigation of how each prediction was computed. As mentioned in the section methods, SHAPely and LIME are used for this purpose.

Figure 6 illustrates the single components which led to this prediction of 6.42 days. The categorical features of this prediction can be seen in the following bullet points:

- discharge_disposition_id_1: Discharged to home

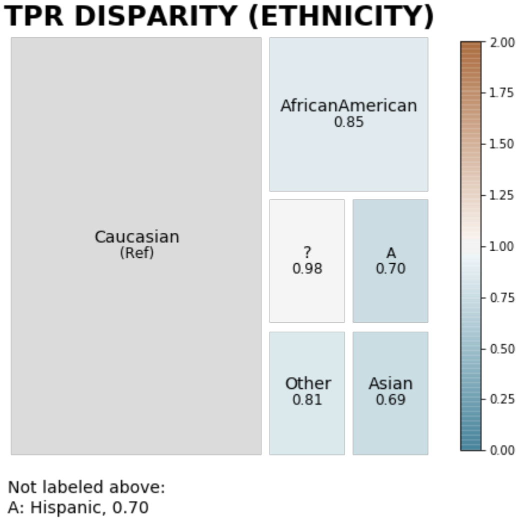

Fig. 4. True positive rate among ethnicities with Caucasiens as reference group with the discrimination boundary at day four

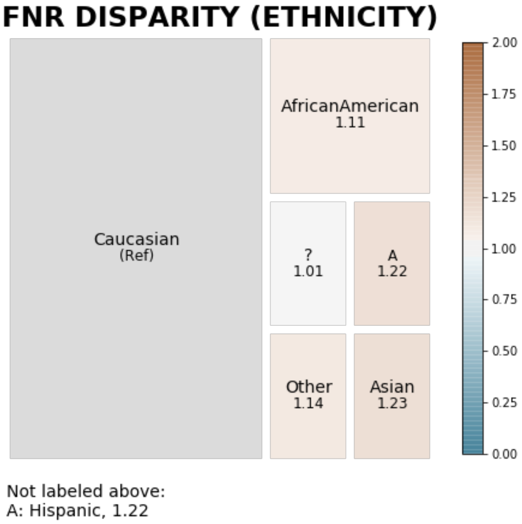

Fig. 5. False negative rate among ethnicities with Caucasiens as reference group with the discrimination boundary at day four

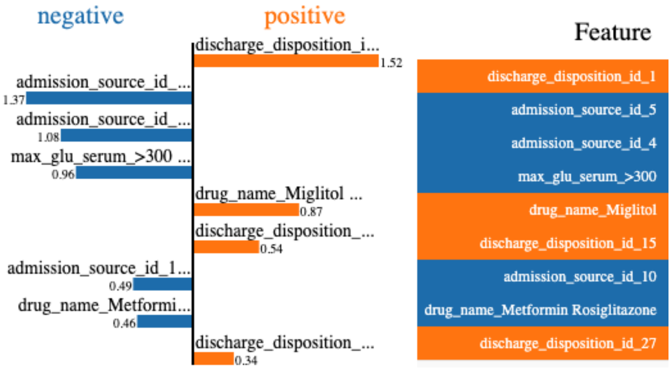

Fig. 6. Important features computed by LIME. The true label for this was 7 days and model predicted a value of 6.42 days.

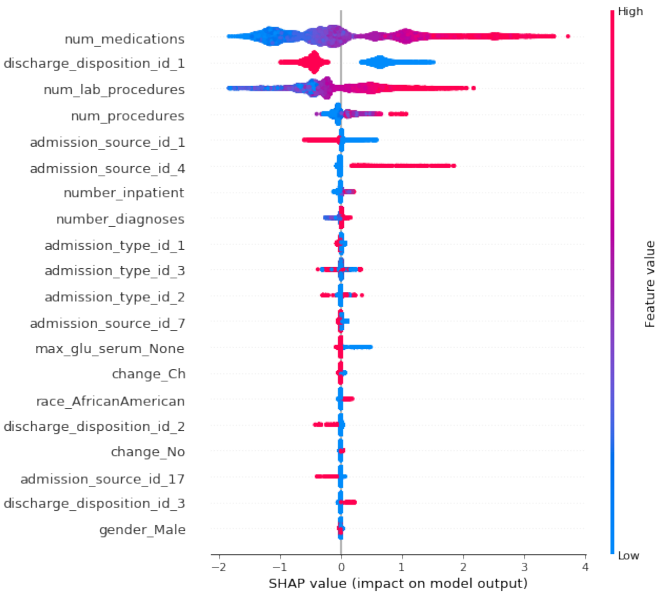

Fig. 7. False negative rate among ethnicities with Caucasiens as reference group with the discrimination boundary at day four

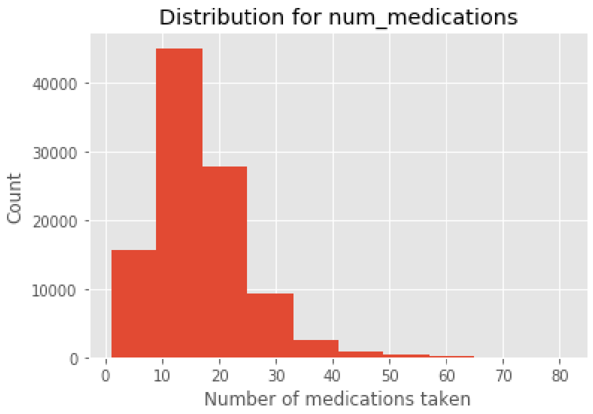

Fig. 8. Distribution of the feature *num_medications*

- 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 an actually **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 *Migitol*, he or she generally will stay up to one day (0.87) longer in hospital than people who don't take this drug. Explanations like this can be conducted for all 36 trained on features for further investigation.

The second interpretational approach are the SHAPely values, which can be seen in Figure 7. This figure lists all features with their distribution on impact for the prediction. For example, the distribution of *num_medications* can be seen in Figure 8. This distribution reveals a right-skewed normal distribution. Based on that distribution, it becomes obvious that a normally distributed numerical feature can have both a negative and positive impact on the prediction. The more medications a person takes, the worse the condition and the longer the hospital stay for this person, 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.

## 4. Discussion

After taking all evaluations and explanatory steps into consideration, can neural networks be applied in the Industry 4.0? There exist many pros and cons.

**Pros**:

- Support for doctor's decision
- Patient selection saves a lot of time
- Possible new insights

**Cons**:

- Unstructured data
- Can it also be applied for other EHR data?
- Necessary needed data amount is unclear

I would argue that these techniques can be a good decision-support technology for doctors. If there are for example 2000 patients stored in the hospital database, this selection process can already shrink the number of people that could possibly be suitable for receiving the novel treatment. Doctors will make the final decision based on their experience, but they can end up saving a lot of time by looking at already 20 pre-selected patients instead of investigating 2000 patient files for suitable patients.

## 5. Conclusion

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

*Apply it without additional doctor's approval?*

For sure not. What if only one person dies because of a false decision of the neural network's prediction? This is a hot topic and the law and regulatory clinical instances like the FDA in the United States or the Arzneimittelbehörde of Germany comes into force when implementing such technologies in the clinical environment.

*Apply it as a support technology?*

Sure 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 interesting insights.

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

Yes. Using first the F1-Score for evaluating the model's performance and then computing the uncertainty estimation with respect to critical features has an overall impro

## 6. Future Work

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

- Try different XAI methods
- Try different Neural Networks or classifiers with GPU?
- Epistemic Uncertainty with different Neural Network Layers: tfp.layers.DenseVariational

## Acknowledgment

Thanks a lot to Philipp Schlieper from the Machine Learning and Data Analytics Lab for a really good supervising through out my project. I can totally recommend this seminar.

## References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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