Interpretability Analysis of MIMIC-III Dataset: Predicting Mortality in Newborns

Ishita Roy, Shweta Shambhavi

University of Texas at Austin

1 Abstract

Interpretability of ML models is of great importance in the healthcare industry, where we need to have trust in our model before concluding any analysis. Interpretability in this case, could mean whether a healthcare professional can follow a model's decision or not. In this paper, we've analysed a globally faced healthcare problem of infant mortality using the transparent interpretable model such as decision tree and logistic regression. Using the MIMIC III dataset, we compared the interpretable model's performance to the performance of non-transparent/uninterpretable models. Since the curated dataset used for this project is heavily unbalanced for the target class, we have high accuracy for all the models. So here, we have observed each model performance on the AUC score. Following AUC as our final deciding metric, logistic regression outperformed most of the model, even some more complex(uninterpretable)models with an average AUC score of 0.978 and AUC of 0.957 for the expired class of the target variable.

2 Introduction

AI in Healthcare sounds suspiciously charming. AI is not always transparent. A model may be perfect for a problem statement but still could very much be a black-box model, hiding crucial decision-making steps in them. So, we want to approach the healthcare data with a comparably simple model, which offers a clear decision defining stage, and see if such models could be a better approach for the healthcare data.

For this project, we are using the MIMIC-III dataset. It is an open dataset curated by the MIT Lab with 60k de-identified health data[1]. In it the patient information is structured in a clean table format, and is ready to integrate into the project. In the AI-Healthcare field, this dataset continuously works to enable and support many research based AI projects. We want to do the exploratory analysis of the data, leading to the deployment of an interpretable classifier which predicts the mortality rate count for newborns.

In healthcare industry, mortality prediction is an important problem[2] . Accurate predictions can help hospitals better allocate medical resources and can provide targeted care for higher risk patients. WHO states that there are approximately 7000 newborn deaths every day, amounting to 47% of all child deaths under the age of 5-years, up from 40% in 1990. The world has made substantial progress in child survival since 1990. Globally, the number of neonatal deaths declined from 5.0 million in 1990 to 2.4 million in 2019. However, the decline in neonatal mortality from 1990 to 2019 has been slower than that of post-neonatal under-5 mortality[3].

We plan to predict the mortality rate for newborns as this is a field which needs helps and there is a lot of scope for improvement. To improve mortality rate in newborns it is very important that we are able to identify and predict probability of fatal illness in advance so that newborns can be provided with right care and medical assistance and a life can be saved. We want to provide the healthcare professionals a helping hand to help save more lives.

We are going to perform interpretability analysis on various classification models such as Decision Tree, Logistic Regression, Naive Bayes, Random Forest, KNN and Gradient Boosting to predict the mortality rate in newborns and identify the most suitable model.

3 Related Work

3.1 Mortality Rate Prediction

Previously there has been extensive research done on mortality rate prediction. [2] In paper ”Interpretable Patient Mortality Prediction with Multi-value Rule Sets, Nationwide Inpatient Sample (NIS) data is used to predict in hospital mortality rate of patients. They have proposed Multi-vAlue Rule Set (MRS) model for making the predictions and Bayesian framework for learning MRS and used F1 score as an evaluation metric. [4] In paper ”Early hospital mortality prediction using vital signals”, MIMIC-III dataset is used to predict early hospital mortality prediction for critically ill patients in ICUs. To derive insight into the performance of the proposed method, several experiments have been conducted using the well-known clinical dataset named Medical Information Mart for Intensive Care III (MIMIC-III). They have used evaluation metrics such as precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). [5] In the paper ”Interpretable Machine Learning Model for Early Prediction of Mortality in Elderly Patients with Multiple Organ Dysfunction Syndrome (MODS): a Multi center Retrospective Study and Cross Validation” They have used MIMIC-III, eICU-CRD and PLAGH-S datasets for model generation and evaluation. They have used machine learning model XGBoost (extreme Gradient Boosting) with the Shapely Additive explanations method to conduct early and interpretable predictions of patients’ hospital outcome. Evaluation metrics AUC, sensitivity, specificity, F1 and accuracy were used for model comparison.

Our work is different from the above listed papers because we are working on prediction of newborn mortality rate using MIMIC-III[1] dataset. We aim to find out the most useful features to predict mortality rate in newborns using various machine learning models. We are performing interpretable analysis using prediction of machine learning classifiers and compare using accuracy, precision, recall and F1 score evaluation metrics.

3.2 Interpretability Analysis

[6] In this paper ”Using interpretability approaches to update “black-box” clinical prediction models: an external validation study in nephrology”, they discuss the validation results of a machine learning model for the prediction of acute kidney injury in cardiac surgery patients initially developed on the MIMIC-III dataset when applied to an external cohort of an American research hospital. [7] In this paper ”An interpretable risk prediction model for healthcare with pattern attention” they use MIMIC-III dataset and proprietary EHR dataset to predict risk of certain diseases. [8] In this paper ”MIMIC-IF: Interpretability and Fairness Evaluation of Deep Learning Models on MIMIC-IV Dataset”, focus is on MIMIC-IV healthcare dataset, and conducts comprehensive analyses of dataset representation bias as well as interpretability and prediction fairness of deep learning models for in-hospital mortality prediction.

Above listed papers have used interpretability analysis on various machine learning and deep learning models by training them using various healthcare datasets and predicting different things. Our work is different from the above work because we plan to perform interpreability analysis of various machine learning algorithms and their performance in predicting mortality in newborns.

4 Methods

We processed four entities, namely, Patient, Admissions, ICU, and the Diagnosis, from the MIMIC III dataset to analyse the 30-day mortality in the Newborn cases. We joined these four tables to create an entire dataset with 32 different features, some categorical and some quantitative. Furthermore, in the ICU-STAYS table, we filtered on the First_Care column to keep the records of the NICU/ Newborns. After filtering, we have 47066 records with 32 features, which we have used to train our interpretable model to predict mortality in Newborns.

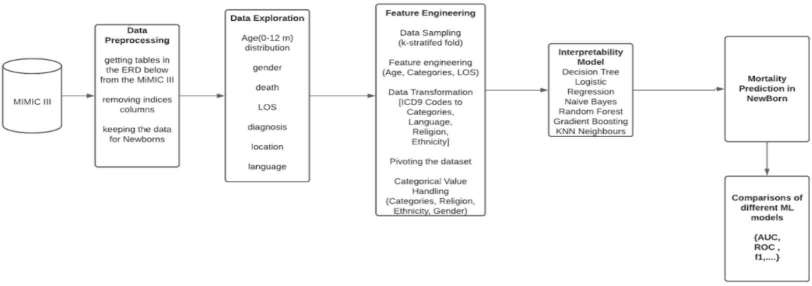
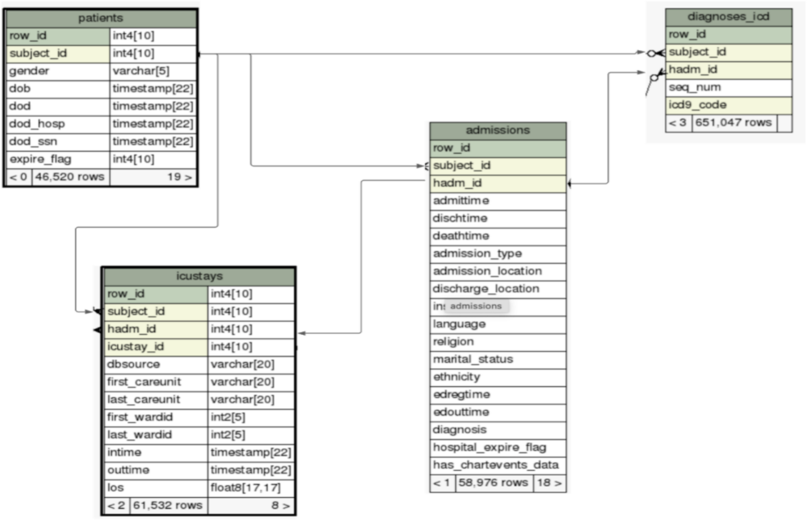

MIMIC III
Data Preprocessing
getting tables in the ERD below from the MiMIC III
removing indices columns
keeping the data for Newborns
Data Exploration
Age(0-12 m) distribution
gender
death
LOS
diagnosis
location
language
Feature Engineering
Data Sampling (k-stratifed fold)
Feature engineering (Age, Categories, LOS)
Data Transformation [ICD9 Codes to Categories, Language, Religion, Ethnicity]
Pivoting the dataset
Categorical Value Handling (Categories, Religion, Ethnicity, Gender)
Interpretability Model
Decision Tree
Logistic Regression
Naive Bayes
Random Forest
Gradient Boosting
KNN Neighbours
Mortality Prediction in NewBorn
Comparisons of different ML models
{AUC, ROC , f1,....}

Fig. 1: Flowchart



patients	
row_id	int4[10]
subject_id	int4[10]
gender	varchar[5]
dob	timestamp[22]
dod	timestamp[22]
dod_hosp	timestamp[22]
dod_ssn	timestamp[22]
expire_flag	int4[10]
< 0 46,520 rows	19 >

diagnoses_icd
row_id
subject_id
hadm_id
seq_num
icd9_code
< 3 651,047 rows

admissions
row_id
subject_id
hadm_id
admittime
dischtime
deathtime
admission_type
admission_location
discharge_location
in: admissions
language
religion
marital_status
ethnicity
edregtime
edouttime
diagnosis
hospital_expire_flag
has_chartevents_data
< 1 58,976 rows 18 >

icustays	
row_id	int4[10]
subject_id	int4[10]
hadm_id	int4[10]
icustay_id	int4[10]
dbsource	varchar[20]
first_careunit	varchar[20]
last_careunit	varchar[20]
first_wardid	int2[5]
last_wardid	int2[5]
intime	timestamp[22]
outtime	timestamp[22]
los	float8[17,17]
< 2 61,532 rows	8 >

Fig. 2: Subset of ERD of MIMIC III Dataset

4.1 Data Exploration

In the Data Exploration step, we have explored different features mainly, Length of Stay(LOS), Gender, Diagnosis, Expire_Flag(Death), Religion, Languages, & Ethnicity. We used different plot options available to us from the Matplotlib and Seaborn libraries to under our data and its potential association with the EXPIRE_FLAG. The dataset is extremely unbalanced with Expired_Flag classes; we have 556 records for class 1(Expired), and the rest 43,410 for class 0. The left-chart in Figure 3. shows the distribution of daily deaths in our dataset, and the right-chart shows the yearly death distribution in the dataset. We have pivoted over the year time frame of the ADMITTIME from the Admission table to get the number of deaths of newborns from the year of 2100 to 2200 (In MIMIC III dataset we have the timeline intact but for privacy concerns the year is moved to range of 2100 to 2200[1]).

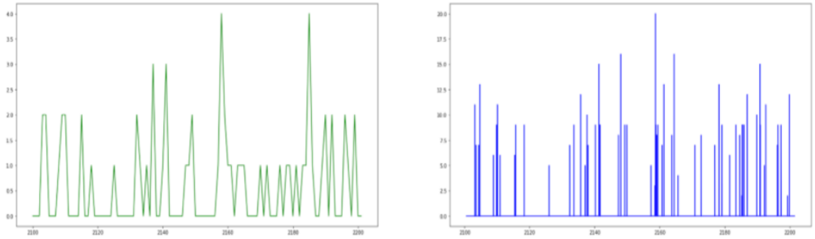
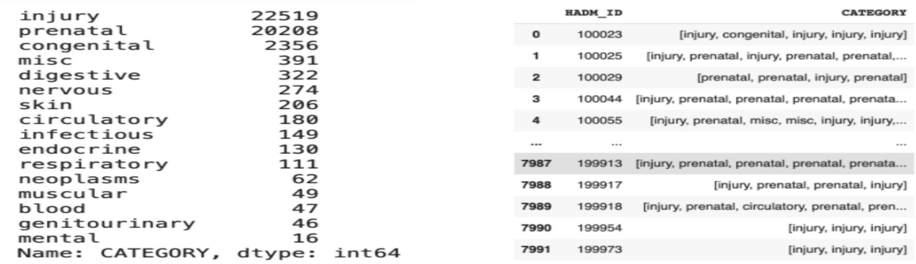


Fig. 3: Daily(left) and Yearly(right) Newborns Death Distribution

4.2 Feature Engineering

The International Classification of Diseases (ICD) is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics[9]. In the MIMIC dataset we have ICD9 Codes, diagnosis and seq-num linked to an admission[1]. We can have multiple ICD9 codes linked to an admission ID explaining the diagnosis of the admission. To group all the different ICD9 codes to an admission, we re-coded a range of ICD9 codes to map to a number ranging from 0 - 17, and linked these re-coded number to various categories like, 'infectious', 'neoplasms','endocrine', 'blood', etc. To have each admission linked with a category as found in Figure 6a. Then finally we grouped HADM_ID on these categories and have a list of multiple categories attached to a distinct HADM_ID in figure 6b. Finally we dummy encode the category feature and summed over any repeated category in the list.



```
injury           22519
prenatal         20208
congenital        2356
misc               391
digestive          322
nervous            274
skin               206
circulatory        180
infectious         149
endocrine          130
respiratory        111
neoplasms           62
muscular            49
blood               47
genitourinary       46
mental              16
Name: CATEGORY, dtype: int64
```

(a) Categories of ICD9

	HADM_ID	CATEGORY
0	100023	[injury, congenital, injury, injury, injury]
1	100025	[injury, prenatal, injury, prenatal, prenatal,...
2	100029	[prenatal, prenatal, injury, prenatal]
3	100044	[injury, prenatal, prenatal, prenatal, prenata...
4	100055	[injury, prenatal, misc, misc, injury, injury,...
...	...	...
7987	199913	[injury, prenatal, prenatal, prenatal, prenata...
7988	199917	[injury, prenatal, prenatal, injury]
7989	199918	[injury, prenatal, circulatory, prenatal, pren...
7990	199954	[injury, injury, injury]
7991	199973	[injury, injury, injury]

(b) Grouping HADM_ID

Fig. 4: Feature Engineering

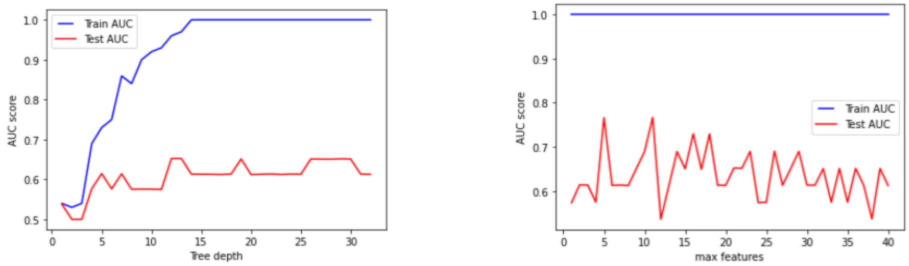
We also cleaned the Religion and Ethnicity values in the dataset and dummy encoded them and some other features such as Gender, Insurance Language, and Admission Location. Lastly we removed the Discharge_Location from the features as it was highly correlated with the Death Flag and was telling of it.

4.3 Machine Learning Models

We have implemented transparent interpretable models such as Decision tree, logistic regression, Naive-Bayes and not so transparent models such as KNN, Random Forest, and the Gradient Boosting classifier to analyse 30-day mortality in the Newborn. For this project our baseline model is 'Decision Tree' considering it's interpretation is the simplest in explaining the relationship between the features as well as with the outcome.

Hypertuning the base decision tree model; we tuned over the max-depths, ccp-alpha, min-sample-splits, min-sample-leaf, and max-features variables of the model. We used the following scale to tune each model parameter and observed overfitting in the case of max-depth and max-features model parameters and rest of the parameters didn't affect the learning of the model. Figures 5a and 5b.

max_depths - np.linspace(1, 32, 32, endpoint=True) ccp_alpha - np.linspace(0.1, 1.0, 10, endpoint=True) Min_samples_leafs - np.linspace(0.1, 0.5, 5, endpoint=True) max_features - list(range(1,X_train.shape[1]))



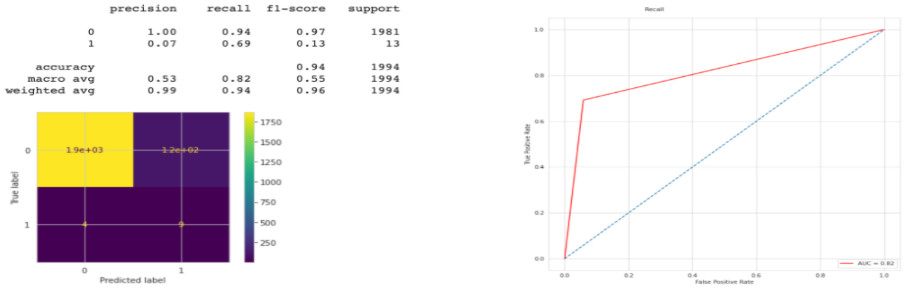
(a) AUC score vs Tree Depth

(b) AUC score vs max features

Fig. 5: Optimising Decision Tree parameters

Entropy criteria performed better with this dataset compared to the Gini. To handle the unbalanced dataset we have used 10 stratified folds and used class_weight parameter as balanced to get the optimised result. Finally we have our baseline decision tree with 0.94 as accuracy, 0.82 as the auc, 0.07 and precision, 0.7 recall and 0.13 and f1-score for the class 1(Expired) of the target feature class.

DecisionTreeClassifier(random_state=42, criterion= 'entropy', max_depth=9, max_features=10, class_weight='balanced')



```
              precision    recall  f1-score   support

           0       1.00      0.94      0.97      1981
           1       0.07      0.69      0.13        13

    accuracy                           0.94      1994
   macro avg       0.53      0.82      0.55      1994
weighted avg       0.99      0.94      0.96      1994
```

(a) True Label vs Predicted Label

(b) True Positive vs False Positive

Fig. 6: Optimised Decision Tree Performance

5 Experimental Design

In this project, we have used the MIMIC III open dataset. It is an open dataset with 26 relational tables in CSV format. It is created by the MIT Laboratory for research advancement in the Healthcare industry and is also available on cloud platforms such as AWS and Google Cloud Platform. We have curated a subset of dataset from 4 of those tables and finally have 7976 rows, 42 columns in our dataset after performing several feature engineering steps. We aim to see how our base model, Decision Tree, performs compared to the Logistic Regression, Naive-Bayes, KNN, Random Forest, and the Gradient Boosting classifiers in predicting the 30- day mortality rate in the Newborn.

Logistic Regression; After hyper-tuning the model, logistic regression was one of the top performers with the highest auc score of 0.978 and accuracy of 0.92. Since the Logistic Regression model is quite transparent and easily interpretable it can be used with or in place of a decision tree if preferred as it's overall model performance is much better than the Decision Tree's performance.

Naive Bayes; Since we have extremely low data points for class 1 and Naive Bayesś uniform distribution assumption for target variable isn't valid here thus the model couldn't perform well for this dataset.

Random forest; As expected it performed better compared to the optimised Decision Tree model using the same optimal parameters of decision tree. Since it is not as interpretable as Decision tree, for cases which require thorough analysis of the outcome it wouldn't be as suitable. Since in this case we are observing the criticality of an admission given the feature set, the model's readability is as important.

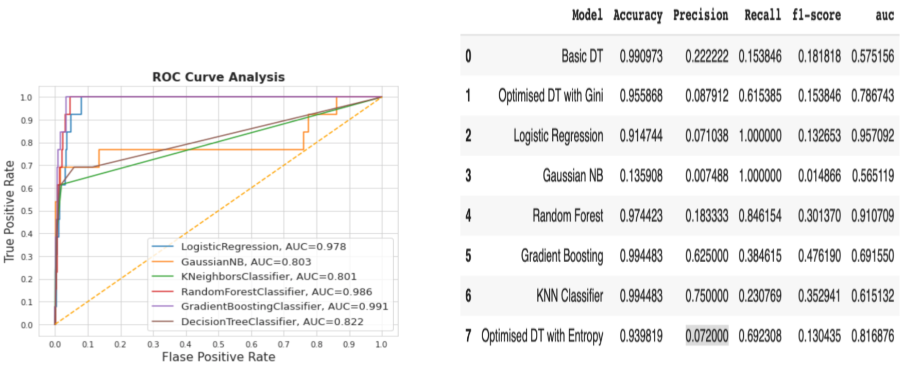
KNN Classifier; KNN had a decent performance with good accuracy and an AUC score. But since it doesn't predict concerning the feature set, even though KNN results are easy to interpret it wouldn't be the best model to understand the importance of the features while predicting for this problem statement.

Gradient-Boosting; Since it is an ensemble of weaker classifiers such as logistic and decision tree, it performed very well and had the avg AUC of 0.991 and AUC of 0.691 for class 1 of the target response.

Finally, we've evaluated the model's performance on the Precision-Recall Curve, AUC curve, accuracy, and f1-score to get the deeper insights of the model's performance and choose the model with the best interest. The purpose of this project is to provide better equipment and inventory estimates for emergency neonatal cases. So, here the false positives would also be acceptable, and we will mainly focus on the AUC score of the model.

6 Experimental Results

We have our optimised Decision Tree model compared with 2 other decision tree models with different model parameters and 5 different transparent and not so transparent models. In interpretable model domain, Logistic regression model outperformed the decision tree and even the uninterpretable models with



	Model	Accuracy	Precision	Recall	f1-score	auc
0	Basic DT	0.990973	0.222222	0.153846	0.181818	0.575156
1	Optimised DT with Gini	0.955868	0.087912	0.615385	0.153846	0.786743
2	Logistic Regression	0.914744	0.071038	1.000000	0.132653	0.957092
3	Gaussian NB	0.135908	0.007488	1.000000	0.014866	0.565119
4	Random Forest	0.974423	0.183333	0.846154	0.301370	0.910709
5	Gradient Boosting	0.994483	0.625000	0.384615	0.476190	0.691550
6	KNN Classifier	0.994483	0.750000	0.230769	0.352941	0.615132
7	Optimised DT with Entropy	0.939819	0.072000	0.692308	0.130435	0.816876

(a) ROC Curve Analysis of various models

(b) Table with values of different metrics for class 1(Expired)

Fig. 7: Experimental Results

average AUC of 0.978 and AUC of 0.91 for class of the target variable. Random Forest and Gradient Boosting also had the similar AUC score but these models have much better f1-score in labelling expired admission cases. We can improve the performance of the model by adding features that will further explain the variance in the model. Since the clinical notes contain much more detail than the tabular field, combining it with tabular data could definitely improve the model's performance, and could be a leading point to proceed the research in the direction of NLP and then introducing the Neural Network Model (black box) and its comparison with the transparent models.

7 Conclusions

This project is on the interpretability analysis of the MIMIC III dataset to predict the 30-day mortality of newborn admission cases. To tag and target critical admissions based on similar past admission records, we implemented a few interpretable and uninterpretable models and analysed their learning of this dataset. We understand such problem statements expect to deliver only the best results while not being a black box. In this project, we explored if simple modelling could work well with healthcare professionals to interpret an admission case. We found that in the case of the two-class classification problem, a simple model with a well-curated dataset and optimal hyper tuned parameters could outperform some of the more complex models(Logistic Regression in this case). We would further like to extend this general idea to the different problem statements as AI can be a helping hand and positive change in the healthcare industry.

References

1. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3** (2016) 160035
2. Wang, T., Allareddy, V., Rampa, S., Allareddy, V.: Interpretable patient mortality prediction with multi-value rule sets. arXiv preprint arXiv:1807.03633 (2018)
3. : Newborns: improving survival and well-being. https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality (September 2020) (Accessed on 04/26/2021).
4. Sadeghi, R., Banerjee, T., Romine, W.: Early hospital mortality prediction using vital signals. Smart Health **9** (2018) 265–274
5. Liu, X., Hu, P., Mao, Z., Kuo, P.C., Li, P., Liu, C., Hu, J., Li, D., Cao, D., Mark, R.G., et al.: Interpretable machine learning model for early prediction of mortality in elderly patients with multiple organ dysfunction syndrome (mods): a multicenter retrospective study and cross validation. arXiv preprint arXiv:2001.10977 (2020)
6. da Cruz, H.F., Pfahringer, B., Martensen, T., Schneider, F., Meyer, A., Böttinger, E., Schapranow, M.P.: Using interpretability approaches to update "black-box" clinical prediction models: an external validation study in nephrology. Artificial Intelligence in Medicine **111** (2021) 101982
7. Kamal, S.A., Yin, C., Qian, B., Zhang, P.: An interpretable risk prediction model for healthcare with pattern attention. BMC Medical Informatics and Decision Making **20**(11) (2020) 1–10
8. Meng, C., Trinh, L., Xu, N., Liu, Y.: Mimic-if: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. arXiv preprint arXiv:2102.06761 (2021)
9. : Icd - icd-9 - international classification of diseases, ninth revision. https://www.cdc.gov/nchs/icd/icd9.htm (Accessed on 05/12/2021).