Int. j. adv. multidisc. res. stud. 2025; 5(1):93-97

Received: 14-11-2024
Accepted: 24-12-2024

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Enhancing Diabetes Hospital Stay Predictions through Interpretability

Frank FH
National Postdoctoral Association (NPA), Rockville MD, United States

DOI: https://doi.org/10.62225/2583049X.2025.5.1.3622
Corresponding Author: **Frank FH**

Abstract
Diabetes Mellitus is a prevalent chronic disease with significant public health implications, often leading to hospitalizations due to complications. Accurate prediction of the length of hospital stay (LOS) is essential for effective patient management and resource allocation. This study focuses on enhancing LOS predictions for diabetes patients using machine learning models, specifically Random Forest and Feed-forward Neural Networks. Utilizing a dataset of over 70,000 patient records from more than 120 U.S. hospitals, key variables such as demographics, admission types, and lab results were analyzed. The findings indicate that incorporating Shapley values improved model interpretability and bolstered confidence in predictive outcomes via enhancing the accuracy and precision of LOS predictions.
CCS Concepts · Machine learning · Healthcare applications · Interpretability in AI

Keywords: Diabetes, Neural Networks, Hospital Stay Prediction, Random Forest

1. Introduction
Diabetes Mellitus (DM) is a globally prevalent chronic disease, exerting considerable pressure on public health systems. According to data from the World Health Organisation, the prevalence of diabetes has been increasing annually and is associated with various complications such as cardiovascular disease, nephropathy, and retinopathy [1]. Hospitalisation is a common treatment modality for diabetes patients, especially when diabetes is accompanied by other complications. The length of hospital stay (LOS) for diabetes patients often reflects the severity of the patient's condition and the therapeutic needs during the hospitalisation period.
Machine learning models could be adapted in healthcare when models are explainable [2]. With the development of medical big data, an increasing number of studies have utilised machine learning, statistical methods, and predictive models to predict the length of hospital stay for diabetes patients. This not only enhances the efficiency of medical resource management but also helps improve clinical outcomes for patients [3]. The main challenge in predicting the duration of hospitalisation for diabetes patients lies in the variability of the patient's condition and the combination of complications, requiring models capable of handling large patient datasets and complex clinical features.

2. Material and Methods
2.1 Prior Investigations & Data
Some studies have employed machine learning algorithms, such as Support Vector Machines (SVM), Random Forest, and Neural Networks, to predict the length of hospital stay [4, 5]. These methods can process a large number of features and perform complex pattern recognition. Research has found that predictive models based on clinical data (e.g., blood glucose control, age, gender, comorbidities) can accurately predict the length of hospital stay for patients. Research Contribution: These studies not only improve the accuracy of predicting the length of hospital stay for diabetes patients but also provide predictive tools for clinicians, helping optimise hospital stay durations and allocate resources reasonably.
Our data derived from the seminal work of Strack *et al*. titled " Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," published in BioMed Research International in 2014 which examines the correlation between the measurement of HbA1c levels and hospital readmission rates among diabetic patients, as elucidated [6]. This research analyzed a substantial clinical dataset encompassing 70,000 patient records, which spans ten

years (1999-2008) and includes data from over 120 US hospitals. Key patient variables such as race, gender, age, admission type, length of hospital stay, medical specialty of the admitting physician, and various laboratory results were scrutinized. The analysis revealed that only 18.4% of patients had their HbA1c levels measured during their hospital stay; however, this subset demonstrated a significant reduction in readmission rates. These findings underscore the critical role of glycemic control in enhancing patient outcomes and underscore the importance of regular HbA1c monitoring in the inpatient setting.

For investigation of relationships between diagnoses and length of hospital stay via machine learning, the initial step involved the de-identification and encryption of all ID fields within the hospital records. Records indicating patient mortality, as identified by the `discharge_disposition_id`, were excluded. Additionally, records containing unknown values in fields unrelated to diagnosis were removed. To enhance computational efficiency, columns with remaining data consisting of a single homogenous value were also eliminated.

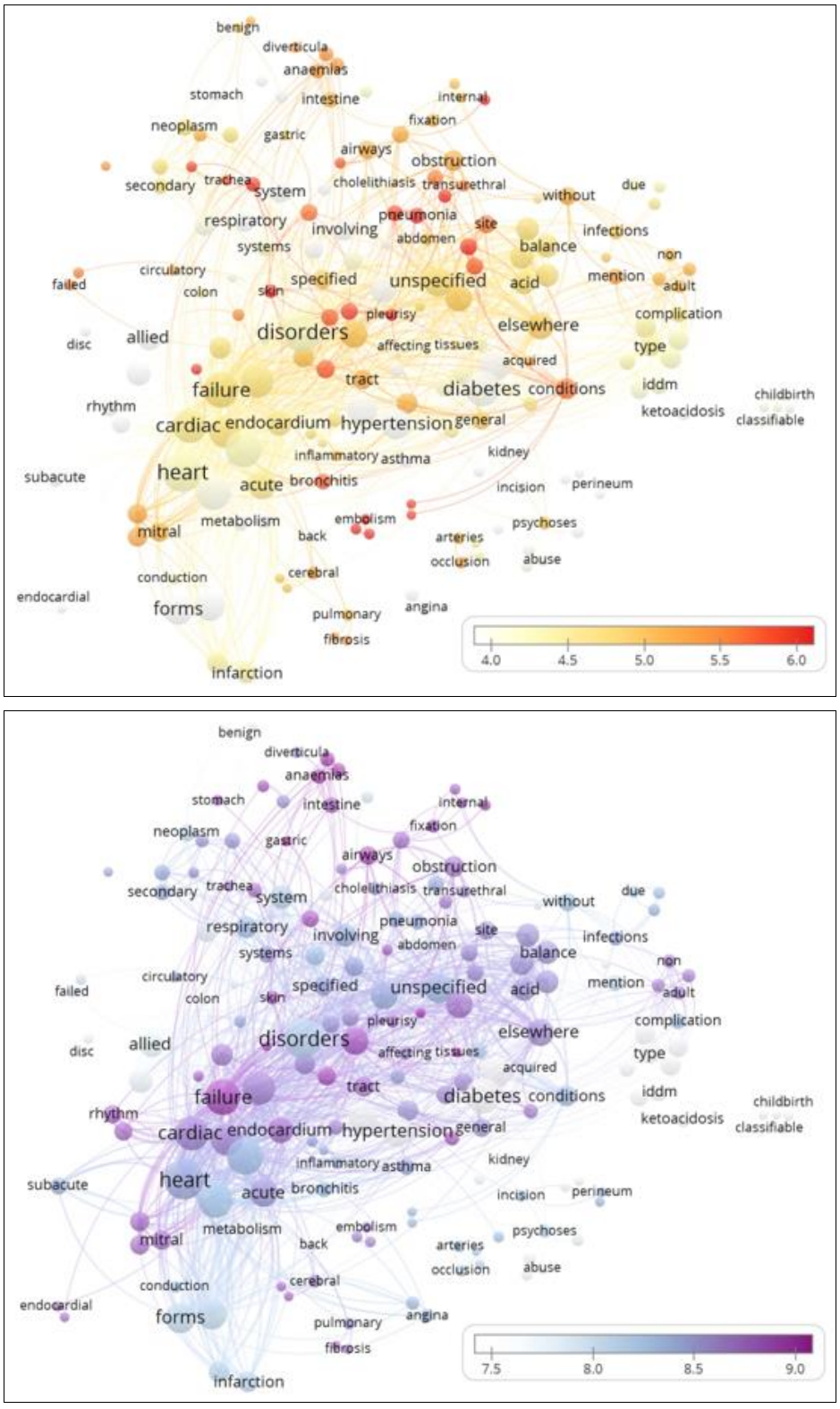

Fig 1: Terminology Word Associations (Occurrence ≥ 7) of Los Versus DX Numbers

Our gradient association maps visualizes the relationships between various ICD9CM based on their co-occurrence in different inpatient admission [7, 8, 9]. Each node represents a word of inpatient ICD9CMs, with the size of the node indicating how frequently the word appears across the dataset. The edges connecting the nodes show the strength of associations between diagnoses terms, with thicker edges representing stronger associations. The color of the nodes reflects the average days of length of stay or average diagnoses number for inpatient records. For example, words with larger, more colorful nodes are both frequently shown in diagnoses fields and with longer inpatient period or with more co-morbidities averagely.

2.2 Methods

2.2.1 Random Forest

Random Forest is an ensemble learning method primarily used for classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs either the mode of the classes (classification) or the mean prediction (regression) of the individual trees [10]. The key features of Random Forest include:

- **Bootstrap Aggregating:** This technique involves creating multiple subsets of the original dataset with replacement. Each subset is used to train a distinct decision tree, thereby reducing variance and improving model robustness.
- **Random Feature Selection:** At each split in the decision tree, a random subset of features is considered, promoting diversity among the trees and mitigating the risk of overfitting.
- **Aggregation:** The final prediction is derived by aggregating the predictions of the ensemble, enhancing overall accuracy and stability.

Random Forests are valued for their flexibility, ease of use, and strong performance across a wide array of data types and domains.

2.2.2 SHAP (SHapley Additive exPlanations)

SHAP is a game-theoretic approach to explain the output of machine learning models. It attributes the prediction of a given instance to the contributions of each feature, ensuring interpretability and transparency. Key aspects of SHAP include:

- **Shapley Values:** Originating from cooperative game theory, Shapley values allocate the "payout" among players based on their contribution to the total gain. In the context of machine learning, these values represent the contribution of each feature to the prediction.
- **Additive Feature Attribution Method:** SHAP falls under this category, where the explanation model is a linear function of binary variables representing the presence or absence of a feature.
- **Consistency and Local Accuracy:** SHAP values satisfy these properties, ensuring that explanations are consistent with the model's behavior and accurate at a local level.

2.2.3 Feed-forward Neural Network

A neural network trained by a back propagation algorithm (multi-layer perceptron) to model complex relationships between inputs and outputs.

The models are trained and evaluated using the following steps:

- Splitting the Data: The dataset is split into training and testing sets, for Feed-forward Neural Network with an 70-30 split.
- Training the Models: Each model is trained on the training set using appropriate hyperparameters.
- Cross-Validation: Cross-validation is used to assess the model's performance and ensure it generalizes well to unseen data.
- Model Evaluation: The models are evaluated on the testing set using metrics such as F-score (F1) to determine their predictive accuracy.

The performance of each model is compared, and the best-performing model is selected based on the evaluation metrics. This model is then used for further analysis and interpretation of the results.

3. Results

3.1 SHAP contributions

Employing the nominal fields diag_1, diag_2, and diag_3, alongside the integer field number_diagnoses, a random forest machine learning analysis is conducted on the time_in_hospital variable. Subsequently, the SHAP (SHapley Additive exPlanations) contribution values are computed for each individual hospitalization record's diagnostic codes and the number of diagnoses. By averaging the SHAP contribution values for each diagnostic code within each field, a ranked table is derived.

To explore whether there are differences in the average contribution values of identical diagnostic codes across different nominal fields (diag_1, diag_2, and diag_3), a selection of 78 diagnostic codes that appear in all three fields was made. These codes were ranked according to their average contribution values across the three fields, resulting in a chart. Additionally, relevant statistical tests were conducted to further analyze the findings.

Table1: Average SHAP Contribution of LOS for Diagnostic Codes Across Nominal Fields

No.	entry	ICD9CM	SHAP	Last	entry	ICD9CM	SHAP
1	diag_2	*heart failure	0.01976	1	diag_1	failed forceps	-0.04073
2	diag_2	*pleurisy	0.01634	2	diag_2	*drug psychoses	-0.03290
3	diag_3	*pleurisy	0.01506	3	diag_2	*chronic ulcer of skin	-0.02399
4	diag_3	*diseases of mitral and aortic valves	0.01331	4	diag_1	limb shortening procedures	-0.02345
5	diag_2	*other cellulitis and abscess	0.00932	5	diag_3	diabetes with unspecified complication, type ii~, uncontrolled	-0.02217
6	diag_1	*acute myocardial infarction	0.00916	6	diag_2	*osteomyelitis,periostitis and other infections involving bone	-0.02054
7	diag_1	pneumonia, organism unspecified	0.00886	7	diag_2	internal fixation of bone without fracture reduction	-0.01943
8	diag_1	*asthma	0.00758	8	diag_1	*disorders of muscle, ligament and fascia	-0.01585
9	diag_2	*essential hypertension	0.00693	9	diag_2	*other venous embolism and thrombosis	-0.01574
10	diag_2	*chronic bronchitis	0.00685	10	diag_1	*benign neoplasm of other parts of digestive system	-0.01450
11	diag_1	*diabetes with renal manifestations	0.00683	11	diag_1	*diverticula of intestine	-0.01398
12	diag_1	*chronic bronchitis	0.00681	12	diag_1	*occlusion and stenosis of precerebral arteries	-0.01362

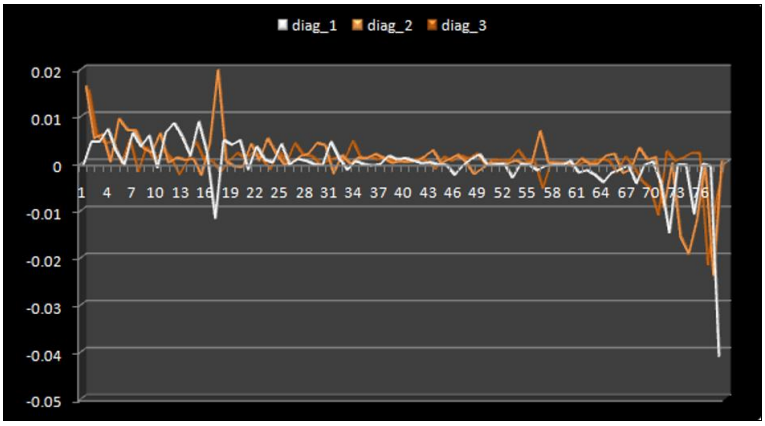

Fig 1: Comparison of Average SHAP Contribution for Identical Diagnostic Codes Across Nominal Fields

3.2 Feed-forward Neural Network models

With all other aforementioned columns processed through data cleansing, the average SHAP contributions of each diagnostic code across the fields diag_1, diag_2, and diag_3 were employed to replace the ICD codes in the Feed-forward Neural Network models, aimed at predicting the quartiles of hospitalization duration. The calculated quartile boundaries for the time_in_hospital variable are as follows: 1-2 days for Quantile 1, 3-4 days for Quantile 2, 5-6 days for Quantile 3, and 7-14 days for Quantile 4. Comparative analyses were then conducted between the modified and original models. The Feed-forward Neural Network models incorporated a 70-30 split and Cross-Validation, resulting in the generation of corresponding confusion matrices. In these NNs, maintain consistency by utilizing conventional configurations: Epsilon of 1.0E-5, 500 training cycles, a learning rate of 0.3, momentum of 0.2, and shuffle normalization during training. These backpropagation algorithm with a multi-layer perceptron architectures were used to establish a single hidden layer of the sigmoid type, with a size equal to (number of columns + 4 quartiles) / 2 + 1.

Table 2: Comparative Analysis of Confusion Matrices for FNN Models Using ICD9CM & AVG(SHAP)

70-30 SPLIT VALIDATION

ICD	true Q1	true Q2	true Q3	true Q4	prec.
pred. Q1	40	35	15	7	41.24%
pred. Q2	35	62	32	30	38.99%
pred. Q3	4	3	3	3	23.08%
pred. Q4	4	14	13	13	29.55%
recall	48.19%	54.39%	4.76%	24.53%	
F1	44%	45%	8%	27%	
accuracy	37.70%				

SHAP	true Q1	true Q2	true Q3	true Q4	prec.
pred. Q1	41	40	17	4	40.20%
pred. Q2	30	48	24	13	41.74%
pred. Q3	8	14	9	8	23.08%
pred. Q4	4	12	13	28	49.12%
recall	49.40%	42.11%	14.29%	52.83%	
F1	44%	42%	18%	51%	
accuracy	40.26%				

CROSS VALIDATION

ICD	true Q1	true Q2	true Q3	true Q4	prec.
pred. Q1	104	86	34	15	43.51%
pred. Q2	146	218	92	94	39.64%
pred. Q3	13	24	15	23	20.00%
pred. Q4	19	49	38	74	41.11%
recall	36.88%	57.82%	8.38%	35.92%	
F1	40%	47%	12%	38%	
accuracy	39.37% +/- 3.88% (MICRO: 39.37%)				

SHAP	true Q1	true Q2	true Q3	true Q4	prec.
pred. Q1	136	107	34	19	45.95%
pred. Q2	95	180	80	67	42.65%
pred. Q3	30	51	28	37	19.18%
pred. Q4	21	39	37	83	46.11%
recall	48.23%	47.75%	15.64%	40.29%	
F1	47%	45%	17%	43%	
accuracy	40.90% +/- 3.46% (MICRO: 40.90%)				

4. Discussion

4.1 SHAP contributions

The analysis of the top 12 and bottom 12 ranked diagnostic codes based on their SHAP contribution values to the time_in_hospital variable reveals several key insights. Notably, cardiac and respiratory conditions, such as heart failure, acute myocardial infarction, and pneumonia, predominantly constitute the top contributors, signifying their association with prolonged hospital stays due to the complexity and intensive care required. Chronic diseases, such as pleurisy and chronic bronchitis, further underscore the extended treatment periods necessary for these ailments. Conversely, the bottom 12 list is primarily composed of minor conditions and surgical procedures, including failed forceps and limb shortening procedures, which typically entail shorter hospital stays, resulting in negative SHAP contribution values. Furthermore, complex diagnoses like diabetes with unspecified complications, type II, uncontrolled, and osteomyelitis, are also associated with negative SHAP values, potentially indicating effective outpatient management or specialized treatments that do not significantly extend hospitalization durations.

The statistical analysis, including both ANOVA and MANOVA tests, did not succeed in rejecting the null hypothesis for the average contribution values of identical diagnostic codes across different nominal fields (diag_1, diag_2, and diag_3). The MANOVA results show that Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root all have p-values 0.8315, indicating a lack of significant differences among the fields. Similarly, the ANOVA results yield an F-value of 0.2833 and a p-value of 0.7535 (>0.05), which further supports the conclusion that there are no significant differences in the average contribution values across the diag_1, diag_2, and diag_3 fields. These findings suggest that the impact of diagnostic codes on the time_in_hospital variable is consistent across the different nominal fields, thereby indicating a uniform influence of the diagnostic codes

irrespective of their classification within diag_1, diag_2, or diag_3. Consequently, these results highlight the stability of the contribution values of diagnostic codes when applied to different fields, implying a relatively uniform effect on hospital stay duration.

4.2 Feed-forward Neural Network models

Based on the analysis of these confusion matrices, it can be observed that the F1 score exhibits an increase in the mean value and a reduction in variance following the substitution of ICD codes with SHAP contributions. This indicates an overall improvement in F1 accuracy, as well as enhanced precision and consistency in the accuracy of the predictions. Although significant differences in the probability distributions of the two paired groups can be determined using Chi-Square tests, when pooling F1 values for the Wilcoxon Signed-Rank Test, the null hypothesis cannot be rejected since the 2-tailed critical value of 4 at a 0.05 significance level is less than the test value of 6. This indicates insufficient evidence to conclude a significant difference in the median of the paired observations.

5. Conclusion

This study has confirmed that using average SHAP contributions to replace ICD codes in Feed-forward Neural Network models enhances both the accuracy and precision of predictions for hospitalization duration. The incorporation of SHAP values has further augmented model interpretability by clearly delineating feature contributions, leveraging Shapley values from cooperative game theory to maintain transparency. This research addresses the significant issue of machine learning models being perceived as "black boxes" by explaining model decision-making processes through interpretable mechanisms such as SHAP. By elucidating the factors that influence predictive outcomes, this work emphasizes the importance of transparency and comprehensibility in deploying machine learning models within healthcare settings, fostering trust and facilitating evidence-based decision-making.

6. Acknowledgment

Humbly would like to express my sincere gratitude to my supervisor for guidance and support throughout this project. I would also like to thank my family and friends for their unwavering encouragement and understanding during this challenging time.

7. References

1. (N.d.). Retrieved 15 November 2024, from: https://www.who.int/news-room/fact-sheets/detail/diabetes.
2. Tseng S, Frank H. Characteristics of inpatient fever: A case in teaching hospital. 2016 International Conference on Computational Science and Computational Intelligence (CSCI). Presented at the 2016 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, December, 2016. Doi: 10.1109/csci.2016.0031
3. Jain R, Singh M, Rao AR, Garg R. Predicting hospital length of stay using machine learning on a large open health dataset. BMC Health Services Research. 2024; 24(1):860. Doi: 10.1186/s12913-024-11238-y
4. Yang CC, Bamodu OA, Chan L, Chen JH, Hong CT, Huang YT, *et al*. Risk factor identification and prediction models for prolonged length of stay in hospital after acute ischemic stroke using artificial neural networks. Frontiers in neurology. 2023; 14:1085178. Doi: https://doi.org/10.3389/fneur.2023.1085178
5. Barsasella D, Gupta S, Malwade S, Aminin Susanti Y, Tirmadi B, Mutamakin A, *et al*. Predicting length of stay and mortality among hospitalized patients with type 2 diabetes mellitus and hypertension. International Journal of Medical Informatics. 2021; 154:104569. Doi: https://doi.org/10.1016/j.ijmedinf.2021.104569
6. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, *et al*. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed Research International, 2014, 781670. Doi: 10.1155/2014/781670
7. H, F.D-C. Metformin: A Versatile Therapeutic Agent in Various Diseases and Health Conditions. Acta Pharmaceutica Hungarica. 2023; 93:45-52. Doi: 10.33892/aph.2023.93.45-52
8. Frank DC. Effectiveness of Music Therapy in Treating Mental Health Disorders in Psychiatry. Acta Psychologia. 2024; 3(1):1-7. Retrieved from: https://psychologia.pelnus.ac.id/index.php/Psychologia/article/view/31
9. Frank H. Predicting Marital Stability: An Approach for More Characteristics. The Asian Conference on Psychology & the Behavioral Sciences, 2024. Doi: 10.22492/issn.2187-4743.2024.5
10. H, F. Exploring the role of urine analysis in early detection of chronic kidney disease. International Journal of Basic and Applied Science. 2023; 12(1):33-38. https://doi.org/10.35335/ijobas.v12i1.248