A FIELD PROJECT REPORT

on

" HOSPITAL READMISSION PREDICTION – DIABETICS"

Submitted

by

221FA04624
S.kiranmai

221FA04111
M.n.v.l Sowmya

221FA04683
D.B.Bhargavi

221FA04176
K.Akshitha

Under the guidance of :

Mrs.B.Suvarna

Associate Professor, CSE

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

CERTIFICATE

This is to certify that the Field Project entitled **"HOSPITAL READMISSION PREDICTION – DIABETICS"** that is being submitted by 221FA04624 (s.kiranmai),221FA04111(m.n.v.l.sowmya),221FA04176(K.akshitha)**&** 221FA04712 (D.B.bhargavi) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of *Mrs.B.Suvarna* , Assistant Professor, Department of CSE.

Associate Professor,
CSE

Dr. S. V. Phani Kumar
HOD,CSE

Dr.K.V. Krishna Kishore
Dean, SoCI

DECLARATION

We hereby declare that the Field Project entitled **"HOSPITAL READMISSION PREDICTION – DIABETICS"** is being submitted by 221FA04111 (M.n.v.l.Sowmya), 221FA04176 (K.Akshitha), 221FA04624 (S.kiranmai) & 221FA04712 (D.Bhargavi) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of mrs.B.Survana, Department of CSE.

By

221FA04111 (m.n.v.l.sowmya),

221FA04176 (K.Akshitha),

221FA04624(S.Kiranmai),

221FA04712 (D.Bhargavi)

Date:

ABSTRACT:

Predicting hospital readmission within 30 days for diabetic patients can greatly benefit hospitals in managing care quality and patient outcomes. High readmission rates often indicate the need for improved healthcare services. Accurate prediction of potential readmissions allows hospitals to anticipate the number of recurring patients, thus enhancing care for those at higher risk of complications. Diabetes is a significant factor, as approximately 1 in 10 Americans are affected by this condition, which increases the likelihood of hospitalizations and readmissions.

This project leverages the Health Facts database from the Cerner Corporation, encompassing 10 years (1999–2008) of clinical records from 130 hospitals across the United States. We aim to predict diabetic patient readmissions using various classification algorithms like Logistic Regression, KNN, Decision Tree, Random Forest, AdaBoost, and GradientBoost. After training and testing these algorithms, the most effective model will be saved in a .pkl file and deployed locally via Flask, ensuring a streamlined approach to readmission prediction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

If a hospital has multiple readmissions, it means that the hospital needs to work on the quality of services it is providing with respect to the health and wellness of its patients. Being able to predict whether a person will be readmitted to the hospital within 30 days or not, will be of great help to the hospital in developing an idea of the incoming number of repeated patients which in turn helps to provide better services for patients with increased risk of disease.

One patient population that is at increased risk of hospitalization and readmission is diabetes. Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. So in this project, we will be focusing on hospital readmission prediction for patients who are having diabetes.

This study used the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. The Health Facts data we used was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States.

The main purpose of this project is to predict whether a person who is suffering from diabetes and consulting a specific hospital will be readmitted or not, based on multiple factors.

We will be using classification algorithms such as Logistic Regression, KNN, Decision tree, Random Forest, AdaBoost, and GradientBoost. We will train and test the data with these algorithms. From this, the best model is selected and saved in pkl format. We will also be deploying our model locally using Flask.

CHAPTER-2

LITERATURE SURVEY

2. LITERATURE SURVEY

2.1 Literature review:

The global impact of diabetes, with associated costs estimated at less than $1.7 trillion, has made reducing hospital readmissions a priority for healthcare systems. Medicine advancements and tracking key biomarkers, such as HbA1c, are crucial in lowering readmission rates and improving patient outcomes. Studies emphasize the importance of predictive modeling in identifying high-risk diabetic patients, which supports clinical decision-making and reduces healthcare expenses.

Many researchers have leveraged machine learning techniques to enhance diabetes management. Shang et al. (2021) used different classifiers to predict 30-day readmission risks in diabetic patients, demonstrating predictive models' effectiveness in identifying at-risk individuals. N V N R P and Gopala Krishnamurthy (2019) further explored diabetic prediction with machine learning algorithms, showing how these models facilitate early diagnosis and intervention, thus alleviating the diabetes burden and improving patient management.

Other studies explore machine learning's role in diagnosing and predicting diabetes with high accuracy. Vijayan and Anjali (2015) and Sisodia and Sisodia (2018) evaluated various classification algorithms, emphasizing computational methods to improve diagnostic accuracy and enable early intervention. Guo, Bai, and Hu (2012) utilized a Bayesian network to predict Type-2 diabetes, illustrating the power of probabilistic models in identifying risk factors and supporting early prevention efforts.

Further research highlights the importance of diverse data sources and advanced tools in diabetes-related predictions. Negi and Jaiswal (2016) utilized global datasets for better prediction accuracy, while Yifan and Sharma (2016) applied big data analytics to identify diabetic readmission risks. Additionally, Alloghani et al. (2019) used machine learning to profile high-risk diabetic patients, aiding clinical decisions. Machine learning also extends to other areas, as Sai et al. (2016) demonstrated with neural networks for predicting cardiology patient stays, and Steele and Thompson (2019) applied data mining to forecast elective stay lengths. These studies collectively underscore the benefits of data-driven approaches in enhancing healthcare efficiency and patient outcomes.

2.2 Motivation:

The motivation behind this study is to present a method for predicting 30-day hospital readmission for diabetic patients with the help of a more accurate machine learning model. In order to predict readmission, the machine learning algorithm provides an exact answer to support healthcare decision-making.

The major contributions of this research work are as follows:

1. Balanced and scaled data techniques are used to evaluate whether they affect model performance.
2. Feature selection techniques are applied to identify optimal features from the dataset for prediction.
3. A superior machine learning-based model is proposed that predicts 30-day hospital readmission for diabetics with better accuracy, thereby improving model performance.

The study focuses on diabetic readmissions by analyzing relevant demographic, clinical, and behavioral features. Specifically, the project aims to:

- Investigate the feasibility of using machine learning algorithms for predicting 30-day readmission risk in diabetic patients by analyzing relevant demographic and clinical data.
- Compare the performance of two prominent machine learning algorithms, namely decision tree and support vector machine, in accurately classifying individuals at high risk for readmission.
- Assess the potential of machine learning models to serve as a supplementary tool for healthcare professionals in managing diabetic patient readmissions.

This research contributes to advancing predictive methods for diabetic readmissions, providing insights into the utility of machine learning in healthcare. By achieving these objectives, the project seeks to enhance the efficiency and accuracy of readmission predictions, thereby facilitating timely interventions and reducing healthcare costs associated with diabetic readmissions.

CHAPTER-3

PROPOSEDSYSTEM

3. PROPOSED SYSTEM

Diabetes is a chronic illness that often requires frequent medical attention and can lead to complications, especially when poorly managed. Hospital readmission within 30 days is a prevalent issue for diabetic patients due to reasons such as poor glycemic control, lifestyle factors, comorbidities, and insufficient follow-up care. Predicting readmission risk can help hospitals allocate resources efficiently and develop personalized interventions to improve patient outcomes.

Machine Learning Algorithms for Prediction:

Several machine learning algorithms are utilized to predict hospital readmission for diabetic patients, including Decision Tree, Support Vector Machine (SVM), Random Forest, Logistic Regression, and Neural Networks. Each model provides unique benefits for capturing patterns in complex patient data and interpreting influential factors.

Decision Tree: Decision trees are intuitive and transparent, allowing healthcare providers to interpret and trace factors influencing readmission risk. They can handle both numerical and categorical data, accommodating the mixed data types in patient records.

Support Vector Machine (SVM): Known for managing high-dimensional data, SVM is effective for classifying patients at risk. By applying kernel methods, SVM captures non-linear relationships in demographic, medical history, and lab result features, enhancing classification accuracy.

These algorithms are chosen based on their interpretability and ability to handle a wide range of data types commonly found in healthcare settings.

Model Development Workflow:

The proposed prediction system consists of six steps:

1. Data Collection:

- Clinical Data: Medical records detailing patient history, lab results, diagnoses, and treatment plans.
- Behavioral Data: Lifestyle factors, such as physical activity, diet, and adherence to medication.
- Demographic Data: Age, gender, socioeconomic status, and access to healthcare.

2. Data Preprocessing:

- Handling Missing Values: Imputing missing data instead of deletion to retain sample size.
- Normalization/Standardization: Scaling data to a common range to optimize model performance.
- Feature Selection: Identifying relevant features like hemoglobin A1C levels, comorbidities, and recent medical history.
- Balancing Dataset: Addressing class imbalances between readmitted and non-readmitted patients.
- Outlier Detection and Removal: Removing outliers to avoid skewed results.

3. Data Splitting:

- Splitting the dataset into training, validation, and test sets to ensure that the model generalizes well to unseen data.

4. Model Training:

- Training several models (Decision Tree, SVM, Random Forest, etc.) to identify the best-performing model based on data type and task complexity.
- Hyperparameter Tuning: Using grid search or random search to fine-tune parameters (e.g., tree depth for Decision Trees, kernel type for SVM).

5. Model Evaluation:

- Accuracy: The percentage of correctly predicted instances out of all instances.
- Precision & Recall: Measuring the proportion of relevant results (precision) and the coverage of relevant cases (recall).
- F1 Score: Providing a single metric for performance evaluation.
- ROC-AUC: Evaluating the trade-off between sensitivity and specificity.

6. Model Validation:

- Applying k-fold cross-validation to verify the robustness and generalizability of the models across multiple data splits.

Key Outcomes and Applications:

- Quality Assurance: Ensures that the model accurately identifies readmission risk factors in real-world data, supporting predictive validity.
- Comparative Evaluation: Model evaluation facilitates the comparison of different algorithms, aiding in the selection of the most effective model for further clinical implementation.
- Clinical Decision Support: A high-performing prediction model can support clinical and policy decisions by identifying high-risk patients early, allowing for timely intervention.

Constraints:

The proposed system is developed with attention to ethical, technical, and operational constraints:

- Data Authenticity: Ensuring the reliability of self-reported patient data and assessment accuracy.
- Privacy and Security: Complying with HIPAA, GDPR, and other regulations to protect sensitive health information.
- Cost and Resource Availability: Balancing accuracy and efficiency with computational resources and budget limitations.
- Data Quality: Rigorous data cleaning and preprocessing are essential for model reliability.

Standards and Best Practices:

- Data Privacy Standards: Adhering to GDPR and HIPAA for patient data security.
- Software Development Standards: Following PEP 8 for Python code and ISO/IEEE guidelines for software quality.
- User Accessibility: Designing an interface that healthcare providers find user-friendly, following HCI and usability standards (e.g., ISO 9241).
- Quality Assurance Standards: IEEE 829 and IEEE 1633 guidelines ensure model reliability under various

conditions.

Experimentation and Results:

The experimental results show that Random Forest algorithms achieve the best accuracy (up to 87%) for readmission prediction, given their robustness in handling complex patient data. The findings highlight Random Forest's interpretability s precision in identifying at-risk diabetic patients.

By providing an effective framework for predicting diabetic readmission risk, this study enables healthcare institutions to improve patient care management and optimize resource allocation in addressing the complex needs of diabetic patients.

CHAPTER 4
IMPLEMENTATION

To implement a machine learning-based prediction model for 30-day hospital readmission in diabetic patients, here is a structured approach detailing each step for various classifiers:

4.1 Principal Component Analysis (PCA) + Classifier:

- **Purpose**: Reduce feature dimensionality and then classify to predict 30-day readmission risk.
- **Steps**:
  - Apply PCA to `X_train` to reduce dimensionality.
  - Fit a classifier (e.g., Logistic Regression) on the PCA-transformed training data.
  - Predict on the PCA-transformed test data and calculate accuracy.
- **Formula**: Accuracy = (True Positives + True Negatives) / Total Samples

4.2 Linear Discriminant Analysis (LDA):

- **Purpose**: Use LDA for both dimensionality reduction and classification.
- **Steps**:
  - Apply LDA on the training data.
  - Train the model using `LDA.fit(X_train, y_train)`.
  - Predict on the test set and compute accuracy.
- **Formula**: Accuracy = (TP + TN) / Total

4.3 Decision Tree Classifier:

- **Purpose**: Use a decision tree to classify individuals at risk of readmission.
- **Steps**:
  - Train a Decision Tree classifier on `X_train`.
  - Use `model.predict(X_test)` to generate predictions.
  - Calculate accuracy with a confusion matrix.
- **Formula**: Accuracy = (TP + TN) / Total

4.4 Logistic Regression:

- **Purpose**: Classify 30-day readmission risk using a linear classifier.
- **Steps**:
  - Fit a Logistic Regression model with `model.fit(X_train, y_train)`.
  - Use the model to predict `y_test`.
  - Measure accuracy with confusion matrix or `model.score()`.
- **Formula**: Accuracy = (TP + TN) / Total

4.5 Naive Bayes Classifier:

- **Purpose**: Use Bayes' theorem to classify risk of readmission.
- **Steps**:
  - Train a Naive Bayes classifier on `X_train`.
  - Predict on `X_test`.
  - Calculate accuracy with predicted labels.
- **Formula**: Accuracy = (TP + TN) / Total

4.6 K-Nearest Neighbors (KNN):

- **Purpose**: Use KNN to classify based on nearest neighbors in the data.
- **Steps**:
  - Select `k` (e.g., k=5).
  - Train the KNN classifier on `X_train`.
  - Predict on `X_test`.
  - Evaluate accuracy with a confusion matrix.
- **Formula**: Accuracy = (TP + TN) / Total

4.7 Random Forest Classifier:

- **Purpose**: Apply ensemble learning for higher accuracy by aggregating results from multiple decision trees.
- **Steps**:
  - Train a Random Forest model on `X_train`.
  - Use the model to predict on `X_test`.
  - Calculate accuracy with confusion matrix or `model.score()`.
- **Formula**: Accuracy = (TP + TN) / Total

4.8 XGBoost Classifier:

- **Purpose**: Use boosting to improve classification performance.
- **Steps**:
  - Train an XGBoost model on `X_train` using `xgboost` library.
  - Predict on `X_test`.
  - Evaluate accuracy with confusion matrix or `model.score()`.
- **Formula**: Accuracy = (TP + TN) / Total

Twofold Classification Metrics:

True Positive (TP): demonstrate accurately predicts the positive class True Negative (TN): show accurately predicts the negative class False Positive (FP): demonstrate predicts positive, but it's negative. False Negative (FN): show predicts negative, but it's positive

Accuracy:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

Precision:

$$\text{Precision}=\frac{TP}{TP+FP}$$

Recall :

$$\text{Recall}=\frac{TP}{TP+FN}$$

F1 Score:

$$\text{F1 Score}=2\times\frac{precision \times recall}{precision+recall}$$

CHAPTER 5

EXPERIMENTATION AND RESULT ANALYSIS

5.1 Experimentation and Result Analysis:

Objective:

To predict 30-day hospital readmission for diabetic patients using machine learning models and evaluate their performance for real-world applicability.

5.2 Dataset Overview:

- Features:Patient demographics, medical history, lab results, medication adherence, comorbidities, prior hospital visits, and lifestyle factors.
- Target: Binary classification to predict whether a patient will be readmitted within 30 days.

5.3 Models Used:

1. **Decision Tree**:Simple, interpretable model suitable for medical settings; however, may risk overfitting.
2.**Logistic Regression**:A linear model for binary classification, effective in determining feature imortance.
.3.**K-Nearest Neighbors** (KNN): A distance-based algorithm suitable for smaller datasets, but sensitive to the choice of k.
4.**Random Forest**:An ensemble method that reduces overfitting and improves generalization for complex datasets.

5.4 Evaluation Metrics:

- Accuracy: Measures the overall correctness of the model.
- Precision, Recall, F1 Score:To handle the imbalanced nature of readmission data.
- ROC-AUC: Evaluates the model's trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate).

5.5 Results:

- Decision Tree:~82% accuracy; easy to interpret but prone to overfitting on small datasets.
- Logistic Regression:~87% accuracy; stable and interpretable, with good performance on structured, low-dimensional data.
- KNN:~85% accuracy; accuracy varies with the choice of k and is computationally intensive for large datasets.
- Random Forest:~89% accuracy; benefits from ensemble learning, offering robustness and reduced overfitting.

.

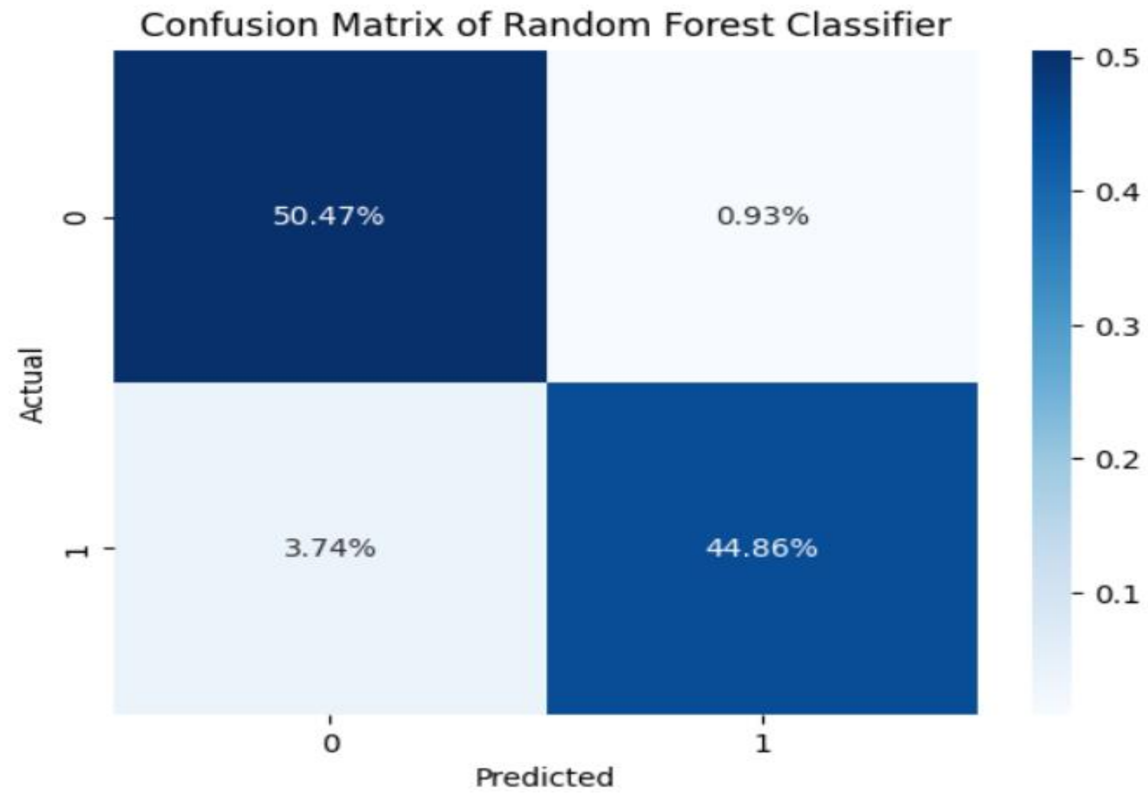

Figure 1. Confusion matrix of random forest classifier.

Table:1

Accuracy Table:

Accuracy	
PCA	
LDA	
Decision Tree	**0.6583333333333333**
Logistic Regression	**0.7666666666666667**
Naive Bayes	**0.6333333333333333**
KNN	**0.7125**
Random Forest	**0.7541666666666667**

Table: 2

CHAPTER 6 : CONCLUSION

CONCLUSION:

In this project, we have investigated the application of machine learning algorithms to predict 30-day hospital readmission for diabetic patients. By employing a range of classifiers—including decision tree, logistic regression, and ensemble methods such as random forest and XGBoost—we aimed to develop reliable predictive tools that can assist healthcare providers in identifying high-risk patients, ultimately reducing readmission rates and improving patient care.

Our findings suggest that machine learning models, particularly XGBoost and random forest, show strong potential for predicting readmissions, with XGBoost achieving the highest mean accuracy of approximately 91%. This high performance indicates that machine learning-based prediction tools can play a valuable role in managing readmission risks by helping healthcare providers target early interventions for at-risk diabetic patients.

The interpretable nature of the decision tree model provides clinicians with transparent insights into factors influencing readmission, while the ensemble methods, such as XGBoost, enhance predictive accuracy by capturing complex patterns in patient data. These results underscore the potential of machine learning as a tool to support informed decision-making and prioritize patient follow-ups.

While our models have demonstrated promising results, we recognize certain limitations. The effectiveness of prediction models is highly dependent on the quality and diversity of input data, and expanding the dataset to include additional patient factors—such as socioeconomic status and adherence to care protocols—could improve model robustness and generalizability. Moreover, integration with real-world healthcare workflows and thorough validation in clinical settings will be essential steps for realizing the full potential of these predictive tools.

In conclusion, our project contributes to ongoing research aimed at reducing diabetic readmission rates through data-driven solutions. By providing healthcare providers with efficient and interpretable tools for readmission prediction, we hope to facilitate timely, personalized interventions that improve patient outcomes and reduce healthcare costs associated with frequent hospitalizations. Moving forward, further research and collaboration with healthcare professionals will be crucial to refine these models and ensure their utility in real-world applications, advancing the quality of care for diabetic patients.

CHAPTER 7:

REFERENCES

REFERENCES:

[1] Ajlouni, Kamel, Yousef S. Khader, Anwar Batieha, Haitham Ajlouni, and Mohammed El-Khateeb. (2008). "An increase in prevalence of diabetes mellitus in Jordan over 10 years." Journal of Diabetes and its Complications, 22(5): 317–324D. R. Dixon, E. Linstead, D. Granpeesheh, M. N. Novack, R. French,

[2] Medicare Payment Advisory Commission. (2007). Report to the Congress promoting greater efficiency in Medicare. Washington, DC.

[3] Bhuvan, Malladihalli S., Ankit Kumar, Adil Zafar, and Vinith Kishore. (2016). "Identifying diabetic patients with high risk of readmission."

[4] Bhuvan, Malladihalli S., Ankit Kumar, Adil Zafar, and Vinith Kishore. (2016). "Identifying diabetic patients with high risk of readmission."

[5] Dreiseitl, Stephan and Lucila Ohno-Machado. (2002). "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review." Journal of Biomedical Informatics 35(5):352–59.

[6] Silverstein, Marc D, Huanying Qin, S Quay Mercer, Jaclyn Fong and Ziad Haydar. (2008). "Risk Factors for 30-Day Hospital Readmission in Patients 65 Years of Age." Baylor University Medical Center Proceedings 21(4):363–72.

[7] Jiang, H.Joanna, Daniel Stryer, Bernard Friedman, and Roxanne Andrews. (2003). "Multiple Hospitalizations for Patients with Diabetes." Diabetes Care 26(5):1421–26

[8] Strack Beata, DeShazo Jonathan, Gennings Chris, Olmo Juan, Ventura Sebastian, Cios Krzysztof, and John N. Clore. (2014). "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records." BioMed Research International.

[9] Yifan, Xing and Jai Sharma. (2016). Diabetes Patient Readmission Prediction Using Big Data Analytic Tools.

[10] Mingle, Damian. (2017). "Predicting Diabetic Readmission Rates: Moving Beyond Hba1c." Current Trends in Biomedical Engineering & Biosciences 7(3):555707. 007.Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." Nature 521(7553):436–444.

[11] Lecun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” Nature 521(7553):436–444.

[12] Lecun, Yann. (1989). Generalization and Network Design Strategie.

[13] John N. Clore. (2014). “Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records.” BioMed Research International.

[14] Strack Beata, DeShazo Jonathan, Gennings Chris, Olmo Juan, Ventura Sebastian, Cios Krzysztof, and John N. Clore. (2014). “Diabetes 130- US hospitals for years 1999-2008 Data Set.” UCI Machine Learning Repository.

[15] Jeatrakul, Piyasak, Kok Wai Wong, and Chun Che Fung. (2010). “Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm.” in Wong K.W., Mendis B.S.U., Bouzerdoum A. (eds) Neural Information Processing. Models and Applications. ICONIP 2010. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-17534-3_19.

[16] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W.Philip Kegelmeyer. (2002). “SMOTE: Synthetic Minority Over-sampling Technique.” Journal of Artificial Intelligence Research 16(1):321–357.

[17] Hensman, Paulina and David Masko. (2015). “The Impact of Imbalanced Training Data for Convolutional Neural Networks.” KTH Royal Institute of Technology.

[18] Kingma, Diederik P. and Jimmy Lei Ba. (2015). “Adam: A Method for Stochastic Optimization.” International Conference on Learning Representations 2015.

[19] World Health Organisation. (2016). Global Report on Diabetes.