

UNIVERSITA' DEGLI STUDI eCampus



eCAMPUS
UNIVERSITÀ

Facoltà di Ingegneria Corso di Laurea Magistrale in

**INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE Classe delle
lauree magistrali in Ingegneria informatica (LM-32)
Curriculum: Artificial Intelligence**

TITOLO DELLA TESI

Artificial Intelligence and Machine Learning Techniques for the
Characterization and Prediction of Diabetes

Relatrice:

Prof.ssa Ing. Patrizia Vizza, Ph.D

Candidato:

Dott.Ing. Leon Martial Doungala Nzoyem
Matricola :001447064

In Memory and Hope

This work is dedicated to all those affected by diabetes — those living with it and those who have lost their lives to the disease. It also honours their families, carers and loved ones who support them every day. May this work contribute to better care, better management and, one day, a cure for this disease.

Table of Contents

1. Introduction.....	7
1.1 Problem Statement.....	7
1.2 Objectives of the Study.....	7
1.3 Structure of the Thesis.....	8
2. Background and Literature Review.....	9
2.1 Overview of Diabetes Mellitus.....	9
2.1.0 Key Definitions.....	9
2.1.1 Types of Diabetes.....	10
A. Type 1 Diabetes.....	10
B. Type 2 Diabetes.....	11
C. Gestational Diabetes.....	12
D. Other Types of Diabetes.....	14
a) Monogenic Diabetes.....	14
b) Secondary Diabetes.....	15
c) Latent Autoimmune Diabetes in Adults (LADA).....	16
2.1.2 Impact of Diabetes on Global Health.....	17
a. Rising Prevalence and Economic Burden.....	17
b. Complications and Mortality.....	18
c. Impact on Vulnerable Populations.....	19
d. Case Study: Diabetes in Italy.....	19
e. Comparison with Other European Countries.....	20
f. Global Comparisons and Ranking.....	20
g. Global Initiatives and Future Directions.....	22
2.2 Introduction to Artificial Intelligence (AI).....	23
2.2.1 Definition , Key Concepts and Subfields of AI.....	23
2.2.2 Evolution of Artificial Intelligence.....	25
2.2.3 Detailed Explanation of AI Subfields.....	26
A. Machine Learning (ML).....	26
I. Supervised Learning.....	26
1. Linear Regression.....	26
2. Random Forest.....	29
3. Support Vector Machines (SVM).....	30
4. Other Notable Algorithms in Supervised Learning.....	32
II. Unsupervised Learning.....	34
1. K-Means Clustering.....	34
2. Principal Component Analysis (PCA).....	35
3. Other Notable Algorithms in Unsupervised Learning.....	39

IV. Reinforcement Learning.....	39
B. Neural Networks.....	40
a. Brief History of Neural Networks.....	40
b. Biological Inspiration and Neural Networks (ANN).....	40
c. The Perceptron.....	42
d. Types of Neural Networks.....	45
e. Feedforward Neural Networks (FNNs).....	45
f. Convolutional Neural Networks (CNNs).....	47
g. Recurrent Neural Networks (RNNs).....	51
h. Long Short-Term Memory (LSTM) Networks.....	53
C. Deep Learning.....	56
a. Brief History of Neural Networks and Deep Learning.....	56
b. Introduction to Deep Learning.....	58
c. Deep Learning Architectures.....	60
D. Generative AI : Brief overview.....	63
2.2.3 Ensemble Learning Techniques.....	64
2.2.3.1 Introduction to Ensemble Learning.....	64
2.2.3.2 Key Concepts in Ensemble Learning.....	65
1. Weak Learners and Strong Learners.....	65
2. Diversity.....	66
3. Aggregation Methods:.....	66
2.2.3.3 Types of Ensemble Learning Techniques.....	67
1. Bagging (Bootstrap Aggregating).....	67
2. Boosting.....	68
3. Stacking (Stacked Generalisation).....	69
2.3 Machine Learning in Healthcare.....	69
2.3.1 Machine learning applications in disease prediction.....	69
1. Predictive modelling in chronic diseases.....	69
2. Personalised medicine.....	70
3. Risk Stratification.....	70
4. Diagnostic Support.....	71
2.3.2 Case Studies: Machine Learning Models for Diabetes Prediction.....	71
• Case Study 1: Logistic Regression for Diabetes Prediction.....	71
• Case Study 2: Support Vector Machines (SVM) in Diabetes Prediction.....	71
• Case Study 3: Neural Networks for Predicting Diabetes Onset.....	72
2.4 Python Libraries for AI and ML.....	73
2.4.1 Scikit-learn.....	74
2.4.2 TensorFlow and Keras.....	75
2.4.3 PyTorch.....	77
2.4.4 LangChain.....	78

3. Machine Learning Techniques for Diabetes Prediction and characterization.....	80
3.1 Data Collection and Preprocessing.....	81
3.1.0 Sources of Data.....	81
a. Dataset description.....	81
3.1.1 Machine Learning project setup and structure.....	84
a) Project Setup and Structure and Technical Steps.....	85
3.1.2 Data Cleaning.....	90
3.1.2.0 Key Definitions.....	90
3.1.2.1 Handling Missing Values.....	91
1. Import Necessary Libraries.....	91
2. Loading the Dataset.....	92
Begin by loading the dataset into a Pandas DataFrame for analysis.....	92
3. Brief Data Exploration.....	92
4. Identify and Handle Missing Values.....	94
3.1.2.2 Correct Inconsistent Data.....	95
1. Identify Numerical and Categorical Columns.....	95
2. Identify Unique Categories in Categorical Columns.....	95
3. Standardise Categorical Variables.....	96
4. Remove ID Column.....	96
3.1.2.3 Identify and remove duplicates and handle outliers.....	97
5. Identify and Remove Duplicates.....	97
6. Identify Outliers Using the IQR Method.....	98
7. Outliers Visualisation and interpretation.....	100
8. Handle outliers in all features.....	105
3.1.3 Exploratory Data Analysis (EDA).....	110
1. Univariate Analysis.....	110
a) Class and Gender Distribution.....	110
b) Diabetic Age Distribution.....	112
2. Bivariate Analysis.....	114
a) Density Plots of HbA1c, Cholesterol, and BMI by Diabetes Class.....	114
b) Correlation Matrix and Heatmap.....	116
c) Correlation Heatmap.....	117
3. Multivariate Analysis.....	118
a. Mean Laboratory Metrics vs. Age for Diabetic Patients.....	118
b. Distribution Visualisation of Key Health Metrics.....	120
c. Interpretation of Key Health Metrics Across Different Diabetes Classes.....	122
d. Gender Differences in Key Health Metrics.....	123
e. Age and Gender Distribution by Diabetes Class.....	125
f. Age Density by Gender and Diabetes Class.....	127
4. EDA Key Findings and Summary.....	129

3.1 Class Distribution.....	129
3.2. Age-related findings.....	129
3.3 Key Health Metrics.....	129
3.4. Significant Correlations.....	129
3.5. Gender Differences.....	130
3.6. Recommendations for model development.....	130
3.1.5 Data Preprocessing and Normalisation.....	131
3.1.5.1 Data Type Analysis and Categorization.....	131
3.1.5.2 Feature Transformation.....	133
3.1.5.3.1 Addressing Skewness in Data Distributions.....	133
3.1.5.5.1 Encoding Categorical Variables.....	137
a) Converting Categorical Variables to Numerical Format.....	137
b) Separe features and targets.....	139
3.1.5.5.2 Feature Scaling.....	141
a) Normalisation of Features : Using StandardScaler.....	141
3.1.5.5.3 Splitting the Dataset into Training and Test Sets.....	144
3.1.6 Feature Engineering.....	145
3.1.6.2 Creating New Features.....	145
3.1.6.3 Visualization and Interpretation of New Features.....	148
3.1.6.4 Saving the Enhanced Dataset.....	150
3.2.0 Machine Learning: Supervised Learning, Unsupervised Learning, and Reinforcement Learning... 151	
3.2.0.1 Supervised Learning: Learning with Labelled Data.....	151
3.2.0.2 Unsupervised Learning: Discovering Patterns in Unlabeled Data.....	152
3.2.0.3 Reinforcement Learning: Learning Through Rewards and Penalties.....	152
3.2 Machine learning model building and selection for diabete Diabetes Prediction and Characterization.....	154
3.2.1 Addressing Class Imbalance [191].....	155
3.2.1.1. Preprocessing: Addressing class imbalance.....	156
1. Class balancing.....	156
• Step 1: Analysing Class Distribution.....	156
• Step 2: Applying Resampling Techniques.....	158
• Step 3: Saving balanced datasets.....	160
3.2.2 Supervised Learning models building Pipeline for Diabetes Prediction and Characterization	161
3.2.2.1 Introduction.....	161
3.2.2.2 Data Preparation.....	162
3.2.2.3 Data pre-processing (reminder and additional steps).....	162
3.2.2.4 Model Development and Evaluation.....	167
a. Overview of preferred Machine Learning Models.....	167

b. Hyperparameter Tuning and Cross-Validation.....	167
c. Model Training and Evaluation.....	171
d. Checks for overfitting and underfitting.....	174
f. Feature Importance Analysis.....	177
3.2.2.4.1 Results and Analysis.....	178
a. Evaluation Metrics Explained.....	178
b. Models performances Analysis.....	180
c. Overfitting and underfitting Analysis.....	183
d. Confusion Matrix Analysis for XGBoost (best model).....	186
e. Feature Importance Analysis.....	187
3.2.2.4.2 Discussion.....	189
a. Pros and Cons of Supervised learning Models Comparisons table.....	190
3.2.2.4.3 Multilayer Perceptron (MLP) for Diabetes Prediction and Classification.....	191
Introduction.....	191
Key Concepts and Definitions.....	192
Development.....	194
Step 1: Import Libraries and Load Data.....	194
Step 2: Hyperparameter Tuning and Model Building.....	195
Step 3: Model Evaluation on Test Dataset.....	197
Step 4: Visualization of Results.....	198
Step 5: Overfitting Analysis.....	200
Step 6: Confusion matrix.....	201
Step 7: Save Metrics.....	202
3.2.2 Unsupervised Learning Models for Diabetes Characterization.....	203
3.2.2.1 Clustering Techniques.....	203
1. Data Preparation:.....	203
2. Dimensionality Reduction:.....	203
3. Clustering with K-Means:.....	204
3.3 Models comparison and selection.....	209
3.3.3 Performance Metrics Comparison (Accuracy, Precision, Recall, AUC-ROC).....	209
Methodology.....	210
3.3.4 Model selection and conclusion.....	215
4. Generative AI for Data Augmentation and Analysis in Diabetes Prediction and Characterization.....	217
4.1 Introduction to Generative AI - LLMs.....	217
4.1.1 Definition and Techniques.....	218
4.1.2 Applications in Healthcare.....	218
4.2.2 XGBoost for Diabetes Classification (Pickled Model).....	220
4.2.3 LangChain for Natural Language XGboost Predictions Interpretations.....	224
4.3 LangChain and RAG for Diabetes Management.....	229

4.3.1 Understanding RAG, Embedding, Vector Databases, and Semantic Search.....	230
4.3.1.1 Embedding: Capturing Semantic Relationships in Data.....	230
4.3.1.2 Vector Databases: Organizing and Querying Embeddings.....	232
4.3.1.3 Semantic Search: Moving Beyond Keywords.....	233
4.3.1.4 Retrieval-Augmented Generation (RAG).....	233
4.3.2 AI Techniques for Diabetes Prediction and Characterization Using RAG and LangChain.....	235
4.3.2.1 Real-World Application of RAG and LangChain Principles.....	235
4.3.2.2 Breaking Down the RAG Process Using the AI Query and Response.....	239
4.3.2.3 Advanced automations Insights with LangChain and RAG.....	240
4.3.2.4 Implementation Highlights from the Repository.....	240
5. Conclusions.....	243
5.1 Summary of Findings.....	243
5.2 Contributions to the Field.....	245
5.3 Final Thoughts on AI in Diabetes Management.....	246
5.4 Closing Remarks.....	246
6. Acknowledgments.....	247
7.1 Books and Journal Articles.....	249
7.2 Online Resources.....	256
7.3 Code Repositories.....	259

1. Introduction

1.1 Problem Statement

Diabetes is a chronic disease that affects millions of people worldwide. It is on the rise, causing serious health problems, economic challenges and reduced quality of life. According to recent reports, the number of adults with diabetes has quadrupled since 1980, mainly due to an increase in type 2 diabetes [1]. Early detection and good management of diabetes are essential to prevent serious complications such as heart disease, kidney failure and blindness [2]. However, traditional methods of diagnosing diabetes often miss the disease in its early stages, leading to delays in treatment and worse health outcomes [3].

Recently, advances in artificial intelligence (AI) and machine learning (ML) have shown great potential for improving healthcare. These technologies help in early diagnosis, personalised treatment and effective management of chronic diseases such as diabetes. AI and ML can analyse large amounts of data to find patterns that may not be obvious to doctors [4]. Using AI and ML to predict and manage diabetes could significantly improve patient care by providing more accurate, timely and personalised treatments [5].

1.2 Objectives of the Study

- The main objective of this thesis is to explore and develop AI and ML techniques to help predict and understand diabetes. Specifically, this study aims to:
- Build and evaluate machine learning models to predict diabetes based on clinical and demographic data [6].
- Explore the use of generative AI techniques to generate additional data for better model training and improved prediction accuracy [7].
- Use LangChain for natural language processing (NLP) to extract important information from unstructured medical records and improve diabetes management [8].
- Compare the effectiveness of different machine learning models and generative AI techniques to find the best approach for predicting and understanding diabetes [9].

- Add to the existing body of knowledge by showing how AI and ML can be practically applied in healthcare, with a focus on diabetes [\[10\]](#).

1.3 Structure of the Thesis

This thesis is divided into several chapters, each covering a specific part of the research:

- **Chapter 2:** Background and Literature Review - This chapter gives an overview of diabetes, AI, and ML. It also reviews the existing literature on how these technologies are used in healthcare, especially in predicting and managing diabetes [\[11\]](#).
- **Chapter 3:** Machine Learning Techniques for Diabetes Prediction - This chapter discusses the methods used to collect and process data, select and train machine learning models, and evaluate their performance [\[12\]](#).
- **Chapter 4:** Generative AI for Data Augmentation and Analysis - This chapter looks at how generative AI techniques can be used to improve data quality and the accuracy of predictive models [\[7\]](#).
- **Chapter 5:** LangChain for Natural Language Processing in Diabetes Management - This chapter explores how LangChain can be used for tasks like extracting information from medical records related to diabetes [\[8\]](#).
- **Chapter 6:** Experiments and Results - This chapter presents the experimental setup and the results of using machine learning and generative AI models, as well as LangChain implementations [\[13\]](#).
- **Chapter 7:** Conclusions - The final chapter summarises the main findings, discusses the contributions of the study, and suggests recommendations for future work .

5. Conclusions

5.1 Summary of Findings

This thesis examined the nexus of artificial intelligence and diabetes management, offering a comprehensive examination of prediction, characterisation, and management solutions. The study employed a range of methodologies, integrating machine learning, deep learning, generative AI, and natural language processing, which culminated in the development of Elyon, a Python Flask-based web application. Elyon provides a practical tool for healthcare professionals to facilitate effective analysis and management of diabetes.

Key findings include:

1. Extensive Literature Review:

- The text provides a comprehensive and detailed understanding of diabetes mellitus, encompassing its various types, the global prevalence of the disease, and its impact on overall health and well-being.
- The study identified the most significant challenges currently facing those managing diabetes, emphasising the necessity for the implementation of AI-driven solutions.

2. Machine Learning in Diabetes Prediction:

- Supervised learning models such as logistic regression, decision trees and neural networks were used to predict incident diabetes.
- Exploratory data analysis (EDA) identified key health measures such as HbA1c, BMI and cholesterol levels as significant predictors.
- Advanced feature engineering and hyperparameter tuning significantly improved model accuracy and reliability.

3. Generative AI for Data Augmentation:

- Techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) addressed data scarcity by generating synthetic data sets.
- These augmented datasets improved the robustness of predictive models and facilitated better generalisation to unseen data.

4. LangChain and Retrieval-Augmented Generation (RAG):

- LangChain's NLP capabilities enabled the interpretation of complex medical records and user queries.
- RAG was used to combine data retrieval with language generation to provide accurate, contextually relevant insights.

5. Web Application Development: Elyon :

- The Elyon app integrated machine learning models, generative AI and conversational agents.
- It featured intuitive visualisations of patient metrics and interactive chat capabilities, making health data accessible and actionable for clinicians.

5.2 Contributions to the Field

This thesis makes several significant contributions to AI and healthcare research:

1. Comprehensive AI Framework for Diabetes Management:

- Demonstrated the integration of machine learning, generative AI, and NLP into a unified system for disease prediction and patient characterization.
- Showcased the practical application of AI methodologies to tackle real-world healthcare challenges.

2. Enhancements in Data Augmentation:

- Developed frameworks for generating synthetic datasets using GANs and VAEs, addressing common issues such as data scarcity and class imbalance in medical datasets.

3. Exploration of LangChain and RAG in Healthcare:

- Provided an implementation blueprint for using LangChain and RAG in medical decision support systems, highlighting their ability to interpret and respond to user queries in a data-driven manner.

4. User-centred design:

- Developed **Elyon**, a web application that combines AI models with user-centric interfaces to improve accessibility and usability for clinicians.

5. Knowledge Contribution:

- Provided a comprehensive review of diabetes management challenges, current AI techniques and their applications.
- Encouraged further research by making partial implementations available via a GitHub repository, balancing transparency with privacy concerns.

5.3 Final Thoughts on AI in Diabetes Management

challenges of diabetes management. This thesis highlights how AI can contribute to:

- **Enhanced Predictive Accuracy:**
 - Advanced machine learning techniques enable early identification of at-risk individuals, facilitating timely medical interventions.
- **Personalized Healthcare:**
 - AI systems like Elyon analyze patient-specific data to deliver tailored recommendations, promoting precision medicine.
- **Streamlined Decision-Making:**
 - AI-powered tools such as LangChain and RAG reduce the cognitive load on healthcare providers by offering evidence-based insights and efficient information retrieval.
- **Scalable and Sustainable Solutions:**
 - Generative AI models address data limitations, ensuring scalability in predictive healthcare applications.

5.4 Closing Remarks

The development of Elyon clearly shows the strong potential of artificial intelligence in transforming healthcare. By combining supervised learning, generative AI, and natural language processing, this project offers a practical and scalable solution to manage diabetes effectively. As AI technology continues to advance, its role in personalized medicine and decision-making will grow even more important, helping doctors and healthcare providers deliver better care and improve patient outcomes. Elyon not only addresses key challenges in diabetes management but also opens the door to exciting possibilities for future research. This project provides a strong foundation for exploring how advanced AI techniques can further enhance global healthcare systems, potentially as the focus of a PhD study.