FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Anusha Ihalapathirana

EXPLAINABLE ARTIFICIAL INTELLIGENCE TO PREDICT CLINICAL OUTCOMES FOR ADULTS WITH TYPE 1 DIABETES

Master's Thesis
Degree Programme in Computer Science and Engineering
May 2022

Ihalapathirana A. (2022) Explainable Artificial Intelligence to Predict Clinical Outcomes for Adults with Type 1 Diabetes. University of Oulu, Degree Programme in Computer Science and Engineering, 69 p.

ABSTRACT

Type 1 diabetes patients are prone to life-threatening conditions. Severe hypoglycemia and diabetic ketoacidosis are such conditions that often require urgent hospital care. Recently, artificial intelligence (AI) techniques have been used to improve the quality of diabetes care and management. These techniques provide a more comprehensive and better experience for patients and their loved ones. The objective of this study is to implement an AI-based explainable solution to predict possible severe hypoglycemia and diabetic ketoacidosis events in T1D patients within the next 12 months. The initial models in this study were built with baseline factors identified in prior research. However, baseline factors alone did not provide enough information, and the models were improved by introducing more features and separating the population by gender. The final predictive models highlighted some of the baseline factors in the original study when predicting the outcomes. Decision support systems based on machine learning models have become a viable way to enhance patient safety by locating and prioritizing high-risk patients. The final models were used to build a decision support system that facilitates precision medicine by prioritizing the high-risk patient group. Moreover, it helps to potentially reduce medical expenses through more efficient resource management.

Keywords: Type 1 Diabetes, Severe Hypoglycemia, Diabetes Ketoacidosis, Machine Learning, Explainable AI, Decision Support System

TABLE OF CONTENTS

FOREWORD

This thesis was carried out as a part of the HTx Next Generation Health Technology Assessment project, conducted in the Biomimetics and Intelligent Systems Group (BISG) at the Department of Computer Science and Engineering of the Faculty of Information Technology and Electrical Engineering at the University of Oulu, Finland.

First, I would like to take this opportunity to express my sincere gratitude to my supervisors Dr. Pekka Siirtola and Dr. Satu Tamminen for their valuable guidance, encouragement, and support throughout the period. I also thank my colleague at BISG, Gunjan Chandra, for her support and encouragement. Finally, I would like to express my profound gratitude to my family and friends for always being there for me with endless love and support.

Oulu, May 31st, 2022

Anusha Ihalapathirana

LIST OF ABBREVIATIONS AND SYMBOLS

Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic
BMI	Body Mass Index
DKA	Diabetic Ketoacidosis
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GDPR	General Data Protection Regulation
HbA1c	Glycated Hemoglobin
HDL	High Density Lipoprotein
HTA	Health Technology Assessment
IQR	Interquartile Range
KNN	K-nearest neighbour
LDA	Linear discriminant analysis
LDL	Low Density Lipoprotein
LGBM	Light gradient boosted machine
MAR	Missing at random
MCAR	Missing completely at random
ML	Machine Learning
NMAR	Not missing at random
PCA	Principal Component Analysis
ROC	Receiver Operator Characteristic
SH	Severe Hypoglycemia
SMBG	Self-monitored blood glucose
SVM	Support Vector Machine
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
UK	United Kingdom
US	United States
USD	United States Dollar

Symbols

$argmax$	Argument of the maximum
a, b	Arbitrary values
C, c	Class
D	Data partition
e	Exponential
log	Logarithmic

N	Number of data points
P	Transformation matrix
S_b	Between-class variance
S_w	Within-class variance
$S(x)$	Probability
$q1$	25^{th} percentile
$q3$	75^{th} percentile
Σ	Summation
μ	Mean
σ	Standard deviation
x	Data point
v	Distinct value in attribute
V, A	Attribute
λ	Eigenvalues

1. INTRODUCTION

The evolution of Artificial Intelligence (AI) has positively affected a lot of traditional research fields that help to solve complex real-world problems. Health care and biomedical applications are one such major area that focuses on improving human lives. Today, AI supports healthcare professionals in many ways, including helping to make proper clinical decisions, enhancing primary care, providing accurate and fast diagnoses, improving patient care outcomes, and increasing productivity of health care delivery. In particular, over recent years, AI-powered precision medicine has become a popular research area worldwide. Precision medicine refers to medical care that selects appropriate treatment for each individual patient or group of patients based on their variability in environment, genetics, lifestyle, and molecular phenotype [1, 2]. This helps to improve medical treatments, early diagnosis, prevent genetically caused diseases, and reduce health care costs. Precision medicine was first used on cancer to make a diagnosis or plan treatments and is now used with many other diseases including diabetes.

1.1. Background

Diabetes is one of the leading diseases that cause mortality and morbidity in the world. In 2021, about 537 million people in the world lived with diabetes, and the majority of them were from low- and middle-income countries [3]. It is a chronic, metabolic health condition that occurs when the human blood glucose level is higher than normal levels. People with diabetes have a high risk of developing serious health problems, including blindness, infections, amputation, heart diseases, and kidney failure. Diabetes is one of the most expensive chronic conditions where the health care expenditure is four times higher for individuals with diabetes compared to people without diabetes [4]. The prominent medical expenditure components include hospital inpatient, treatments for diabetic complications, diabetes supplies, and physician office visit costs [5].

Type 1 diabetes (T1D) is one of the main types of diabetes. People with T1D need to be closely managed with daily care. They need a healthy lifestyle, daily insulin replacement, and regular blood glucose monitoring to keep their blood glucose level in a healthy range. While there are many complications related to T1D, severe hypoglycemia (SH) and diabetic ketoacidosis (DKA) remain two common, and major acute complications [6]. These two are extreme ends of T1D complications which can cause seizures, coma, brain damage, hypokalemia, and swelling inside the brain. Proper diabetes management and early detection of possible events are important preventive actions to avoid such serious complications.

There is a need for research that focuses on adults with T1D due to the growing population of adult patients and the risk associated with these two complications. Many studies are carried out on the prevention of severe hypoglycemia and diabetic ketoacidosis in children [7]. However, the associated treatments and factors for SH and DKA are different in adults. Furthermore, most of the existing works are carried out to find the risk factors associated with these complications, and the majority of them are based on statistical methods. Currently, the studies on predicting SH and DKA event occurrences in adults are limited. Recently, various researchers focused their work

on predicting severe hypoglycemia event occurrences in adults. These studies have limitations of a small number of features, a small sample size, and limited forecasting windows [8, 9].

1.2. Objectives

The main objective of this study is to implement an Artificial intelligence (AI) based solution to predict the possible risk of having severe hypoglycemia and diabetic ketoacidosis events in adult T1D patients within the next 12 months.

These prediction models will serve as a decision support tool for healthcare personnel to treat T1D patients effectively. Transparency is one of the main principle in AI ethics and model transparency is an essential requirement in medical applications [10]. It is important to have an explanation for the predicted outcome and the contribution of each feature behind the prediction. This tool visualizes and interpret the impact, and feature contributions of the individual predicted result. This assists health care professionals to understand the reasons for each individual's predictions and provide personalized treatments. Furthermore, health care personnel can view major parameters that impact the predictions of the entire group.

The final goal is to use these prediction models to build a decision support system to categorize the patients into high-risk, moderate-risk, and low-risk patient categories. Patients in different categories require different treatments and healthcare checkup procedures. It is important to identify and closely monitor patients in the high-risk category. The boundaries of these categories are dependent on the allocated costs for medical resources, and they can be changed based on the application. This system will facilitate precision medicine to treat at-risk type-1 diabetic patients and help authorities significantly reduce medical expenses through efficient resource management.

Addressing the current research limitations, this study is focused on 7155 individuals in the T1D exchange clinic registry dataset [11], aged 26 to 93 years old with T1D for more than two years. The work of the thesis was carried out with Python language and aimed to build prediction models to detect possible SH and DKA events beforehand.

This research was carried out as part of the HTx Next Generation Health Technology Assessment project [12]. HTx is a Horizon 2020 project supported by the European Union lasting for 5 years from January 2019. The main aim of HTx is to create a framework for the Next Generation Health Technology Assessment (HTA) to support patient-centered, societally oriented, real-time decision-making on access to and reimbursement for health technologies throughout Europe.

The structure of the thesis is as follows. Chapter 2 provides an overview of the related work on T1D, SH, and DKA. Chapter 3 describes in detail the theoretical background of the machine learning steps, methods and tools used in this study, and Chapter 4 presents the experiments carried out and the final results of the study. Chapter 5 contains the discussion of these results, limitations, and future work, while Chapter 6 presents the conclusion of the thesis.

2. RELATED WORK

The first records of diabetes are found in ancient Egyptian papyrus records [13]. It has been described as a medical condition of excessive urination, thirst, and weight loss. Since then, with the evolution of medical science and technology, numerous studies have been carried out to treat and increase the life expectancy of patients with diabetes. This chapter covers related work on diabetes, T1D, and related complications. Section 2.1 briefly describes diabetes mellitus and Section 2.2 covers details about T1D. Sections 2.3 and 2.4 discuss related work on severe hypoglycemia and diabetes ketoacidosis, respectively. The final section covers the research details on the original research paper of the replication study in this thesis.

2.1. Diabetes Mellitus

Diabetes mellitus is a fast-growing global issue with significant health, social and economic consequences. Diabetes occurs when the human blood glucose level is high. Insulin is a hormone produced in the pancreas that helps glucose to get into blood cells and reduces the level of glucose in the blood. Inability to produce enough insulin or to use it effectively leads to an increase in blood glucose levels. High blood glucose levels damage the internal organs in the human body over time. This disease is common among adults and the vast majority of cases are reported in China with about 141 million people [14].

Type 1, type 2, and gestational diabetes are the main types of diabetes [15]. Gestational diabetes is a temporary medical condition that develops in women during their pregnancy period and usually disappears after giving birth. Type 2 diabetes (T2D) is the most common type of diabetes and it occurs when the human body does not make enough insulin, or the body becomes resistant to insulin. Due to this insulin resistance, blood glucose levels increase. Over 90% of people with diabetes have T2D and are usually older adults [16]. However, due to poor diet, obesity, and lack of physical activity, the rate of T2D incidents has lately been increasing among children, adolescents, and younger adults. Currently, there is no cure for T2D, but it can be prevented or delayed by managing a healthy lifestyle [17].

While T2D is more commonly diagnosed later in life, type 1 diabetes (T1D) is often identified early in life, but it can develop at any age. It is not caused by poor diet or unhealthy lifestyle habits; it is a genetic disorder. The cause of T1D is an autoimmune reaction that destroys the pancreas beta cells which are responsible for making insulin [18]. When the pancreas stops producing or produces a small amount of insulin, it increases blood glucose levels. Unlike T2D, T1D is not a disease of affluence.

People who live with diabetes need proper management to maintain their health and quality of life. Treatment and management options depend on the type of diabetes and the lifestyle of the individual person. This study focus on T1D patients. However, regardless of the diabetes type, it is important to manage it effectively to maintain the wellbeing of an individual.

2.2. Type 1 Diabetes

Recent analysis discovers that the incidence and prevalence of type 1 diabetes are significantly increasing in the world [19]. The reasons for these higher rates of incidences are not fully discovered, but studies show that environmental factors like viral infections, hygiene, gut microbiology, obesity, and genetics are contributed to this increment. The T1D incidences are higher in the Nordic countries where the highest rates are found in Finland, while the lowest rates are found in South American and Asian countries [20].

T1D is also known as Juvenile-onset diabetes because it is commonly diagnosed during childhood. T1D patients always depend on insulin, and they need proper management including daily blood glucose monitoring, nutrition, regular exercise, and emotional support. People with T1D face different challenges depending on their age and other social factors. Since most of them are diagnosed in childhood, they spend a majority of their lives with the disease. The longer the diabetic duration, the higher the possibility of getting T1D-related complications.

Insulin therapy has progressively increased the lifespan of T1D patients [21]. However, without proper management, T1D can cause life-threatening complications, including heart attack, neuropathy, nephropathy, diabetic retinopathy, amputation, infections, and pregnancy-related complications which develop over the years. SH and DKA are two common acute complications among T1D patients. Frequences of SH and DKA events are high among adults with long-term diabetes. Preventing these complications will lead to a better quality of life.

T1D is a lifelong disease and associated with a significant direct cost for both patients and medical organizations. Having a family member with T1D will also affect the other members of the family, especially when the patient is a child. Studies reveal that many family members show anxiety regarding the health of their loved ones [22]. This disease heavily impacts their day-to-day life activities which are indirectly connected to the economy. As an example, employees with T1D or parents of T1D children tend to take lots of leaves and this causes productivity loss. Furthermore, it is difficult to calculate the physical and social costs associated with type 1 diabetes. However, Dall et al. [23] estimated the annual cost of diabetes in the United States and discovered approximately 14.9 billion USD is related to T1D and it is 8.6% of the economic burden. This figure includes 10.5 billion USD in medical costs and 4.4 billion USD indirect costs. Similar research showed close figures and found costs related to T1D are disproportionately higher than the number of individuals with T1D when compared to individuals with T2D in the United States [24].

Even though there are huge investments in high-quality medical research and facilities, a cure for T1D is not available yet. Therefore, it is important to support T1D awareness and effective management to ensure long-term health and economic sustainability.

2.3. Severe Hypoglycemia

Severe hypoglycemia is a common complication in people with a long history of T1D. The probability of having SH events is increased with age. A severe hypoglycemia

event can be defined as a diabetic emergency of having a low blood glucose level and requiring assistance from another person to get treatment [25]. Hypoglycemia has three levels, asymptomatic, mild-moderate symptomatic, and severe hypoglycemia [26]. Most people feel symptoms of hypoglycemia at its mild-moderate level including, dizziness, sweating, hunger, and weakness. It is possible to prevent serious problems by treating hypoglycemia at this point. During a severe hypoglycemia event, there is a high risk of having seizures, loss of consciousness, falls/accidents, blurred vision, and inability to follow normal tasks. Hypoglycemia needs immediate treatments including getting blood glucose levels into the normal range. Untreated severe hypoglycemia can be fatal and Jensen et al. [27] have observed a high mortality risk with the people suffering from hypoglycemia events.

HbA1c is a form of hemoglobin that is connected to sugar and it plays a major role in diabetes. To prevent SH events, it is recommended to optimize the glycated hemoglobin (HbA1c) levels in children. However, studies show that this treatment approach is not practical for adults [28, 29]. With the growing interest in T1D, numerous studies are carried out on different paths to find the factors associated with severe hypoglycemia events in adults and children.

2.3.1. Risk Factors Associated with Severe Hypoglycemia

The risk factors associated with SH events vary in different patient categories. Weinstock et al. [30] carried out statistical analysis with adults who were 60+ years old and had T1D for more than 20 years. The study reveals that hypoglycemia unawareness and high glucose variability play a major role in recent SH event occurrences and are less related to lower HbA1c or mean glucose levels. A study carried out on 415 subjects with 4.5 to 6 years of diabetic duration found that better glycemic control and older age increase the risk of having SH events [31]. According to Giorda et al. [32] from 206 subjects, one in six patients had at least one SH event experience in the past 12 months, and severe hypoglycemia incident rates were higher when the patient had a previous history of severe hypoglycemia, neuropathy, long diabetes duration, and on polypharmacy. A Danish-British multicentre survey carried out with adult T1D patients shows lack of awareness, peripheral neuropathy, and smoking were significant risk markers to increase SH occurrences [33]. Moreover, a study carried out by Sämann et al. [34] identifies low BMI, low HbA1C, insulin therapy, and female gender as factors to increase the risk of SH events. This retrospective study focused on patients with diabetes in Germany. These patients were treated for diabetes in primary care and they discovered this primary care helps to achieve good glycemic control with infrequent SH events.

Pregnant women with T1D have a high risk of SH events. Ringholm et al. [35] identify the frequency of SH event occurrences are higher in early pregnancy than in the period before pregnancy. History with severe hypoglycemia, long diabetes duration, low HbA1C in early pregnancy, lack of hypoglycemia awareness, fluctuations in plasma glucose values, and excessive use of insulin injection play major roles to increase the SH event probability during early pregnancy [35].

2.3.2. Severe Hypoglycemia Prediction Models

Predicting the possible risks of SH events can help to prevent future occurrences. Recently, Dave et al. [8] developed a machine learning model, an optimized random forest classifier, for probabilistic prediction of hypoglycemia events in youth with T1D. The model used 26 features including demographic data and glucose pattern data. The model was able to predict hypoglycemia in 30 and 60 minutes time horizons with more than 91% sensitivity and 90% specificity. Zhang et al. [36] developed a classification tree model to predict the occurrences of future hypoglycemia events in a one-hour window. The study used glucose measurements and insulin infusion rate data of 3116 T1D patients in the intensive care unit to provide early warnings of hypoglycemia events in the clinical environment. Georga et al. [37] presented an SVM model to predict possible hypoglycemia events in 15 patients using recent glucose profiles, insulin intake, meals, and physical activities as predictor variables. They were able to achieve a sensitivity of 92% and 96% for 30 min and 60 min prediction horizons, respectively. Lately, Ruan et al. [38] proposed a method to predict the risk of inpatient hypoglycemia, using 17658 inpatients' data including patient demographics, administrated medications, vital signs, and laboratory results variables. Among 18 predictive models, the best results were achieved with the XGBoost model with an AUROC of 0.96.

Severe hypoglycemia is the main limitation to achieving strict control of glucose levels in T1D patients. Several studies focus on predicting possible SH event occurrences in adult patients. Cox et al. [39] developed a sliding algorithm that is able to predict imminent SH events in adults with T1D. They used patterns in self-monitored blood glucose (SMBG) readings and achieved 58-60% prediction with only three preceding SMBG readings. In 2017, Schroeder et al. [40] developed two models to predict six months' risk of hypoglycemia, one model with 16 features and another with six features. The cohort size of 31 674 consisted of both type 1 and type 2 diabetes patients. The performance of the models, presented in c-statistics was 0.84 and 0.81 for the 16 and 6 feature models, respectively. The predictor features include age, diabetes type, HbA1c, history of prior hypoglycemia events, and insulin-related data. They used statistical analysis method, the Cox regression model for the counting process in their work. They also found that individuals with T1D have a higher risk of having SH events compared to T2D patients. Reddy et al. [41] proposed two ML models to predict hypoglycemia events in adults at the start of aerobic exercise, a decision tree-based model and a random forest model. Both models were trained using 154 observations of 43 adults with T1D aged 21 to 45 years old. The decision tree model identified two critical features, heart rate and glucose level at the start of the exercise, that help to predict possible hypoglycemia events during exercise with 79.55% accuracy. The random forest model showed higher accuracy with 86.7% where it used 10 features including sex, BMI, energy expenditure, average daily insulin dosage, and blood glucose value at the start of exercise.

Over the years, the use of machine learning models to predict severe hypoglycemia events has significantly increased. Various research was conducted to forecast possible severe hypoglycemia events in different patient groups using different predictor variables. However, current methods have several common limitations. One of them is a limited prediction window. Detecting the possible risk of having SH events in

the early stage helps patients to deal with SH more appropriately. Most of the above prediction models are able to predict severe hypoglycemia events within a 30 to 60 minute prediction horizon [8, 36, 37].

While there are a lot of risk factors involved in increasing the risk of SH events, blood glucose level-related data is the dominating predictor vector in most of the studies [36, 37, 39]. ML requires a large amount of quality data to provide an accurate and complete outcome. However, it is hard to achieve with clinical trial data and some of these studies explore their work with a small number of data subjects [37, 41] or with statistical models [39, 40]. Some studies have been carried out to predict SH events considering all patient groups in the dataset. But the risk factors across patient groups can be drastically different. For example, children have different risk factors from adults with T1D, so it is important to focus on different health groups when building a prediction model.

Fear of hypoglycemia reduces the quality of life of the patients and their family members [42]. It can be too late to prevent the SH event at the time patient recognizes its symptoms [43]. Hence, it is important to detect the possible risk of severe hypoglycemia events with a proper prediction horizon. Such a model can warn T1D patients about possible future risks and help them to manage their condition appropriately.

2.4. Diabetic Ketoacidosis

Diabetic ketoacidosis is another serious acute metabolic complication of T1D. It develops when the human body lacks insulin and cannot use glucose as an energy source. When there is not enough insulin to consume glucose, the human body starts to burn fat to get energy. This process produces ketone acids in the liver and sends them to the bloodstream [44]. An excessive amount of ketones in the blood lead to diabetic ketoacidosis events that cause hypokalemia, swelling inside the brain, fluid inside the lungs, and kidney damage [45]. DKA occurs most commonly in T1D patients as well it can also occur in T2D patients when they are in stressful situations like during infections or surgery. Infections, heart attacks, strokes, pancreatitis, certain medicines, and alcohol are some of the common triggers for DKA events in diabetes patients.

Usually, DKA symptoms can develop quickly and they include excessive thirst, blurred vision, shortness of breath, abdominal pain, and vomiting. DKA is a medical emergency and when a patient shows symptoms of it, he/she is required to visit the hospital for treatments. Significant morbidity, utilization of healthcare resources, and costs are associated with DKA treatments [46]. Research carried out in 2017 shows that the average time for hospitalization related to one DKA episode was 5.6 days and the average cost for an episode was EUR 2065 per patient [47]. Even though the hospitalization period is short, the direct cost of managing DKA episodes in adults is relatively high. The United Kingdom (UK) National Diabetes audit have shown that nearly 4% of people with T1D in the UK experience DKA incidences each year and about 6% of DKA incidences have occurred in adults newly diagnosed with T1D [48].

2.4.1. Risk Factors Associated with Diabetic Ketoacidosis

Similar to SH studies, various researchers focus on their work to find the risk factors associated with DKA events in T1D patients. However, compared to SH studies, there are a limited number of studies focused on DKA in adults [49]. A systematic literature review carried out by Farsani et al. [49] observed that the frequency of DKA events decreases with the increasing age of the patient. They found that several factors including lower socioeconomic status, female gender, depression, and poor glycaemic control have considerably increased DKA risk in adult patients. A study carried out by Ehrmann et al. [50] identified similar results. Socioeconomic disadvantage, adolescents aged 13 to 25 years, female gender, high HbA1c values, depression or eating disorder, and previous DKA event history reveal strong evidence for increased risk of incidence of DKA in patients with T1D. Al-Obaidi et al. [51] aimed to discover the socioeconomic factors related to DKA in T1D patients. After carrying out statistical analysis with 147 patients, they found that the younger age, underweight, low education level, unemployment, travel, uncontrolled HbA1C, home glucose monitoring count, and termination of insulin are associated with a high risk of DKA. They also found factors that do not have a significant impact on the risk of DKA including gender, marital status, smoking status, income, frequency of HbA1c checking, and family history.

2.4.2. Diabetic Ketoacidosis Prediction Models

Currently, there are a limited number of studies focused on the early prediction of possible incidences of DKA in patients with T1D. Gilhotra et al. [52] conducted research to predict DKA in children with T1D using nasal capnography, venous blood gases, and urinary analysis data. They use Chi-squared tests and receiver operating characteristic analysis methods to analyze the data. Recently, Li et al. [53] evaluated the performance of different ML approaches in identifying DKA risk factors and predicting DKA events. The study focused on 3400 T1D adult patients with DKA occurrences and 11780 with no DKA occurrences. Best results were achieved with the XGBoost classifier with an AUC of 0.887. HbA1c levels, non-DKA hospitalization, white blood count, hemoglobin, pulse rate, and BMI features were identified as the major risk factors.

There are only a few studies on predicting the possibility of having DKA incidences. However, it is important to conduct further research to identify possible risk groups beforehand since DKA occurs in a short period of time.

2.5. Original Research Paper of the Replication Study

Identifying the risk factors that are associated with SH and DKA events is an important research area in type 1 diabetes. Researchers have found that there are a number of risk factors that need to be considered based on the patient groups. One major factor is the socioeconomic background of the patient, as it can often have a significant impact on

diabetic management. One of the aims of this thesis is to replicate the study carried out by Weinstock et al. [54] with machine learning techniques.

Weinstock et al. [54] used the T1D exchange registry dataset to discover the factors associated with SH and DKA in adults with T1D in the United States. The cohort included 7012 subjects filtered based on the age of 26 to 93 years old and having T1D for more than two years. This study is based on the hypothesis that the rates of DKA are positively associated with HbA1c, SH is inversely associated with HbA1c, and frequencies of having SH and DKA are associated with socioeconomic status, diabetes duration, age, and pump usage. Occurrences of one or more SH and DKA events in the previous 12 months were considered for their study. They used logistic regression models to evaluate the association between selected factors and event occurrences in the last 12 months. Both univariate models and multivariate models were developed.

This study mainly focused on 13 factors, Age, T1D duration, Mean HbA1c in the past year, Gender, Race/Ethnicity, Household annual income, Insurance status, Education, BMI, Insulin delivery method, Insulin, 1st degree relative with T1D, and current smoker. They tried to find the association of these features with the frequency of SH and DKA occurrences.

2.5.1. Risk Factors of Severe Hypoglycemia in Adults

In the cohort, 4973 patient data were available for SH events, and from that 587 participants reported one or more SH events in the previous 12 months. Their research discovered that high SH frequency can be seen with longer diabetic duration, HbA1c values less than 7.0% or higher than 7.5%, and low insulin requirements. Moreover, the frequency of SH events was higher among the participants with low education levels, low household income, and no private insurance.

In univariate models, they found that higher SH event frequencies are associated with older age, non-Hispanic black/Hispanic ethnic groups, participants who use injections as an insulin method, and participants who smoke at the time of data collection. But these three factors did not show a significant impact on the multivariate model.

These results reveal the importance of diabetes management for T1D patients since a higher number of SH events are strongly associated with individuals who achieve the HbA1c goal of 7.0% or have higher HbA1c levels (> 7.5%). The authors point out that achieving optimal glycemic control itself is a major obstacle to reducing SH risk and health care personnel should consider modifying HbA1c goals, especially the individuals with very low HbA1c levels and previous SH history.

2.5.2. Risk Factors of Diabetic Ketoacidosis in Adults

DKA event occurrences data were available for 6796 patients and from that 326 patients have reported one or more DKA event occurrences in the past 12 months. The authors discovered that the frequency of DKA event occurrences was lower with the increasing age. HbA1c levels, female gender, lower socioeconomic status including low education, low income, and no private insurance showed outstanding association with higher DKA frequencies in the multivariate model. However, the frequency

of DKA was not associated with the diabetic duration. Similar to SH, participants with non-Hispanic black/Hispanic ethnicity and current smokers showed higher DKA frequencies in univariate models, but not in the multivariate model. Also, DKA was not significantly associated with the insulin administration method.

Weinstock et al. [54] used statistical methods to analyze and find the associations of risk factors with SH and DKA occurrences. In this thesis, machine learning models were used on the same dataset used by Weinstock et al. to predict the possible risks of SH and DKA occurrences in the next 12 months. Initial steps were carried out with the features mentioned in [54] for the model predictors.

3. METHODOLOGY

Real-world medical data provides an opportunity to understand diseases and conditions better. It also helps to solve challenging healthcare problems using machine learning techniques. This chapter briefly introduces the background of the theoretical topics used in the thesis work. Section 3.1 introduces the data set used to build predictive models. Section 3.2 covers the theoretical background of machine learning processes and techniques, and Section 3.3 describes the ML model performance evaluation techniques. Section 3.4 presents the tool used to explain the output of machine learning models.

3.1. Dataset: T1D Exchange Registry

The reports of national diabetes statistics show that approximately 1.87 million Americans live with T1D by 2020, and it is estimated that 64,000 people are diagnosed with T1D each year [55]. T1D Exchange network is a nonprofit clinical research established in 2009. It aims to fill the knowledge gaps on the optimal approaches to managing type 1 diabetes in the United States (US). The T1D Exchange consists of a clinical data registry, a patient-centric website, and a biobank. The registry is a longitudinal, prospective data repository that collects core clinical and laboratory data of T1D patients with the collaboration of 67 clinical centers in the US [11].

There are ethical concerns associated with collecting and using medical data. This study was carried out in compliance with the ethical principles of the Declaration of Helsinki and the standards of good clinical practice. Individuals with T1D diagnosed and receiving care from the clinical center in the T1D Exchange network were eligible for the study. Written informed consent was obtained from all eligible adult participants and consent was obtained from the parents / guardians of the children. Baseline data were collected at enrollment and once a year during follow-up visits.

This rich data set consists of patients' longitudinal data from 2010 to 2018. It is publicly available for research purposes, and all patient identification and dates have been de-identified. The registry contains 493 attributes that were collected using database history, physical exam forms, laboratory test results, HbA1c test results, and medical conditions. Participant questionnaires were used to collect diabetes history, management, complications, monitoring, quality of life information, socioeconomic factors, general health, family history, pregnancy, and menstrual data. Data from 25759 individuals were available in the study from ages less than one year to 93 years. 50% of the population were female, 82% were white non-Hispanic, and 90.6% were under 26 years of age when diagnosed with T1D.

Figure 1 illustrates the T1D duration of participants. The average duration of T1D was 11.43 years at the enrollment, and among the participants with available data (n = 25759), 2464 subjects had T1D for more than 30 years and 1082 for more than 40 years. The mean age at diagnosis and the mean age at consent was 11.6 and 23.0, respectively. 21530 subjects (83.5% of population size) were diagnosed with T1D at less than 18 years of age, and 41 subjects were diagnosed at 65 years of age or older (Figure 2). This is expected behavior since the diagnosis of T1D is more common during childhood. The mean HbA1c level was 8.3%, and as described in Figure 3,

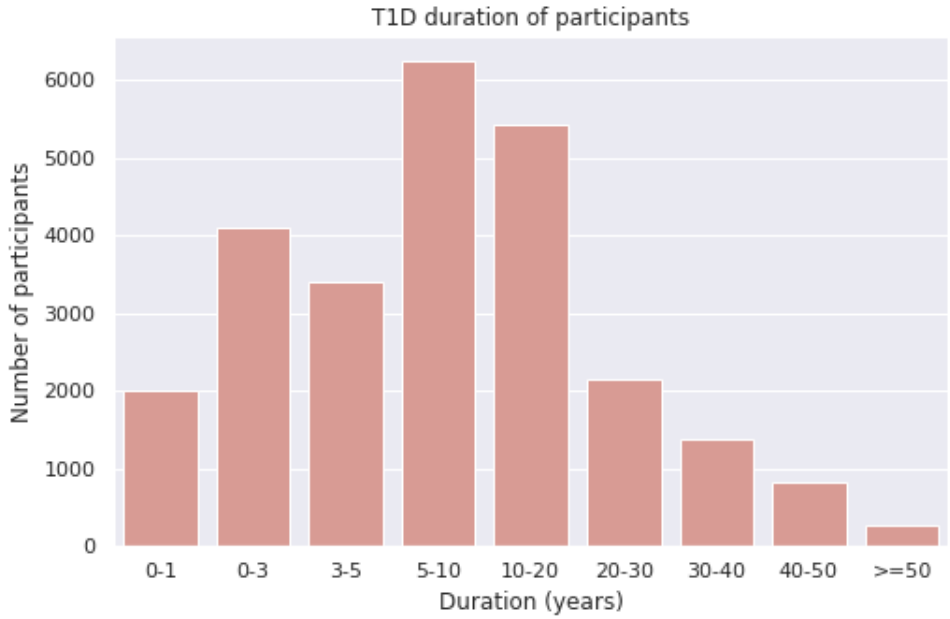

Figure 1. T1D duration.

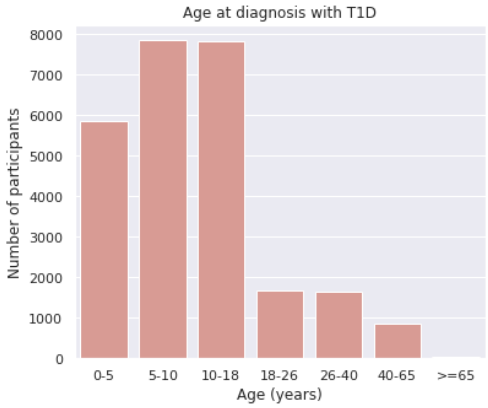

Figure 2. Diagnosed Age.

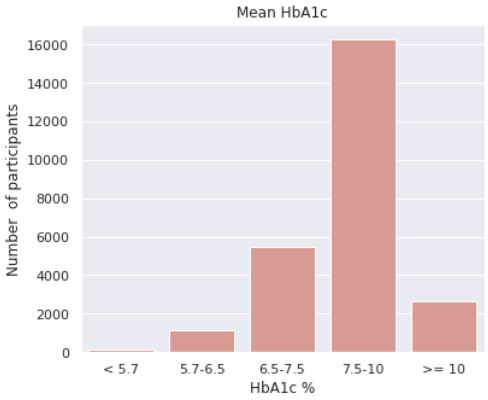

Figure 3. Mean HbA1c levels.

1098 participants had a mean HbA1c level below 6.5%, and 2856 had a level above 10.0%.

519 participants reported at least one severe hypoglycemia event in the prior 12 months, and 1309 reported DKA occurrences. Figure 4 shows the number of participants who had occurrences of SH events in the previous 12 months with their corresponding mean HbA1c levels (average HbA1c over 12 months). Similarly, Figure 5 illustrates the patients with at least one DKA event occurrence in the prior 12 months. Both figures show a similar trend, and a higher frequency of SH and DKA can be observed in patients with HbA1c levels in the 8.0-9.0% range.

This large data set provides real-world information covering a wide range of age, ethnic/racial, and socioeconomic groups of individuals with T1D. This dataset can be used to understand the disease and develop health care systems to improve patient-centered care.

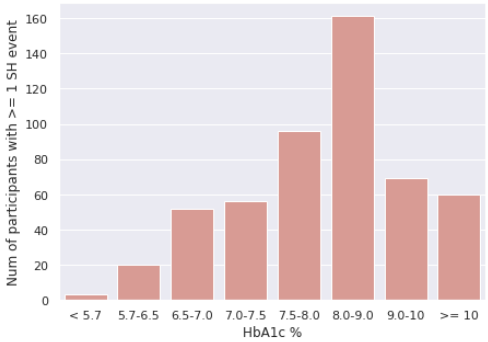

Figure 4. SH events and HbA1c%.

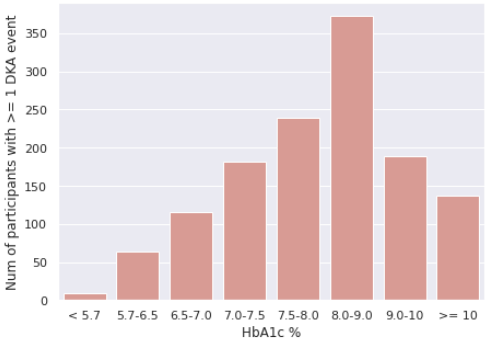

Figure 5. DKA events and HbA1c%.

3.2. Machine Learning

Machine learning is a rapidly growing research area in AI. Machine learning algorithms use historical data to find hidden patterns and use them to make decisions. Nowadays, ML techniques solve different problems in various medical domains, including medical diagnosis and prognosis. ML provides effective and accurate solutions to a wide range of complex tasks like detecting possible medical conditions, disease predictions, forecasting possible health risks, assisting with healthcare records, drug development, therapy support, patient management, and many more. Healthcare data contains a large amount of complex data that provides valuable information about a patient's condition. Proper data management, analysis, and machine learning techniques are required to obtain meaningful information from these data.

Machine learning is classified into four categories according to its type of learning: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

Supervised learning: This technique uses a labeled dataset to train the model and uses that trained model to classify or predict unseen values. Supervised learning is divided into two categories; regression and classification. The regression model predicts the output as a real value, for instance, weather forecasting, while the classification model provides categorical output such as spam detection [56].

Unsupervised learning: Data labeling is an expensive and laborious task, and it may require domain knowledge. Unsupervised techniques are used to find hidden patterns in the data without prior knowledge of the labels. This technique can explore and identify the structural patterns in the data. Clustering is the most common technique in unsupervised learning, where it discovers groups (clusters) in the dataset based on the data distribution [56].

Semi-supervised learning: Most real-world problems fall into this category, where some of the data are labeled in a large dataset. Semi-supervised learning combines previous techniques that use both labeled and unlabeled data during the training period [57].

Reinforcement learning: Nowadays, reinforcement learning is used to solve lots of real-world problems, including industrial robotics, fraud detection, and

autonomous driving. In this method, the agent learns to reach the goal by performing actions [58]. The agent receives positive or negative feedback based on the action it takes, which helps to achieve its long-term goal.

Figure 6 depicts the general machine learning process, and the theoretical background of these steps is discussed in sub-sections.

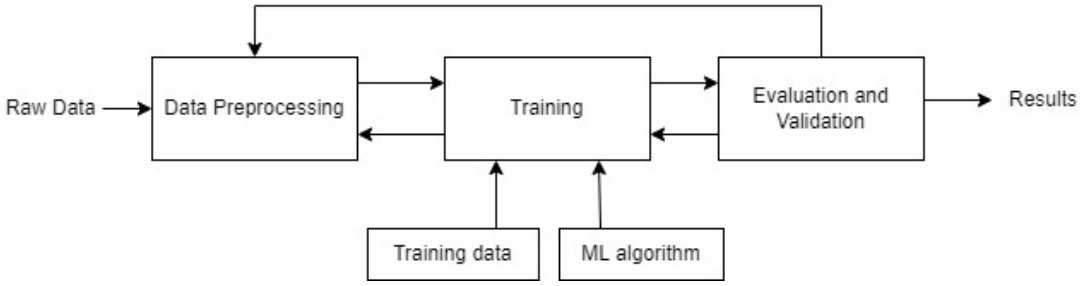

Figure 6. Machine Learning Process.

3.2.1. Pre-Processing

Data quality is one of the main factors to consider when building a successful ML model. It is well-known that quality data provides quality outputs in ML. Although real-world data produce opportunities to address challenging issues, they often come with poor quality. It is difficult for an ML model to discover hidden patterns if the dataset contains irrelevant, unreliable, noisy, or redundant data [59]. It is important to prepare the data before using it to conduct an effective analysis. The steps involved in data pre-processing are data integration, cleaning, transformation, and data reduction. Figure 7 presents the common data pre-processing steps.

Data integration: Real-world data come from different heterogeneous sources. Data integration is an essential pre-processing step for medical data since information is collected from patients, clinical records, medical history, pharmacy, and test results. Collected data can be stored in several databases, record cubes, or flat documents. These different data sources are combined to produce unified data during this step. Data integration helps to provide a uniform view of data while preserving its information.

Data cleaning: It is expected to have noisy samples in real-world data, and there are many possibilities of having duplicated or mislabeled data after the data integration step. Inaccurate data provides unreliable information. Data cleaning is the process of handling noisy, incomplete, irrelevant, duplicated, or missing data in a data set. Some common cleansing steps are removing duplicate and irrelevant samples, handling structural errors, removing outliers, and handling missing values. However, these may vary depending on the data set.

Removing all unwanted samples, including irrelevant and duplicate data, is important to improve the training speed of the ML model. Irrelevant data are the observations that do not fit into the problem. Large datasets have inconsistent

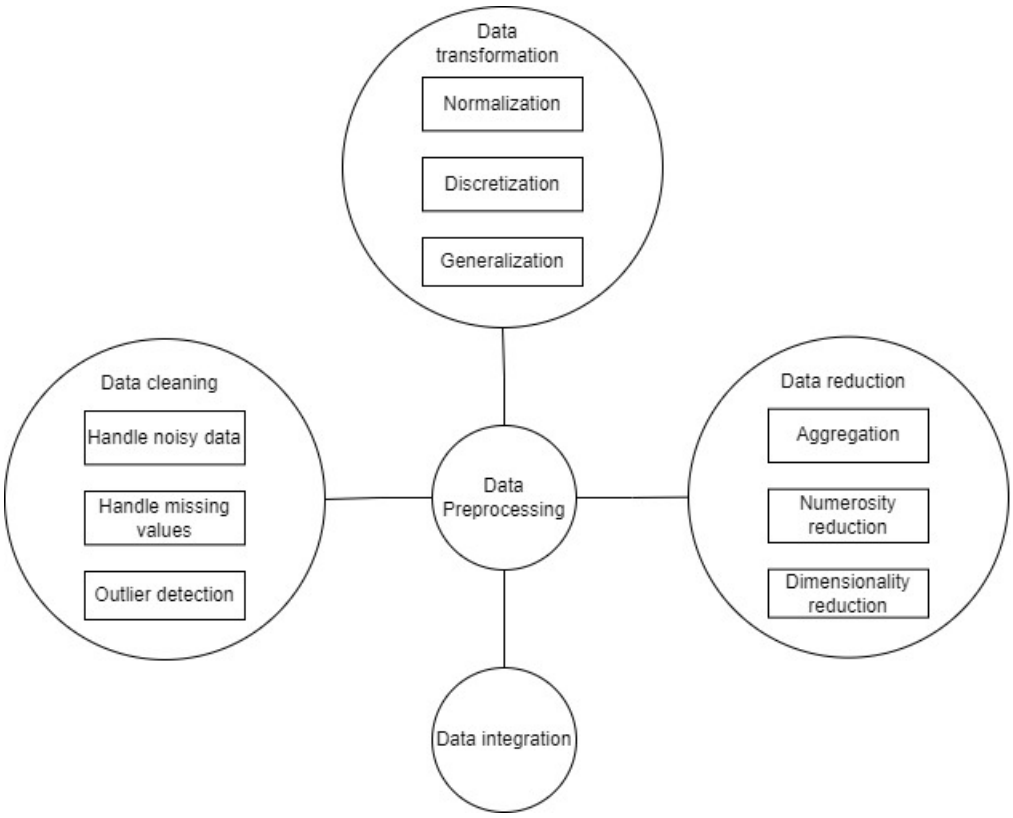

Figure 7. Data pre-processing.

data, typos, different formats, and naming conventions. For example, the date of birth can be found in both formats 'MM / DD / YY' and 'DD / MM / YY'. These structural errors must be fixed during the pre-processing stage.

An outlier is a sample that deviates from the other samples in the population [60]. Detecting outliers in the dataset may reveal bad data, but outliers caused by natural events may provide interesting information about the dataset in some scenarios. However, the presence of outliers adds significant uncertainty to the outcome; therefore, it is necessary to remove outliers to have an accurate model [61]. There are multiple ways to detect outliers in a dataset. Box plots are a visual method that uses the interquartile range (IQR) to detect outliers. IQR is the difference between the 75^{th} percentile (q3) and 25^{th} percentile (q1). Box plots use this method to highlight the outliers by computing upper and lower bounds as:

$$Lowerbound = q1 - \frac{3}{2} * IQR \tag{1}$$

$$Upperbound = q3 + \frac{3}{2} * IQR \tag{2}$$

In this method, the points above and below these bounds are considered outliers. However,this outlier removal method does not consider how the data is distributed and may work poorly with a skewed data distribution.

The standard score (Z-score) is another method for outlier detection when the data follows Gaussian distribution [62]. Z-score helps to understand how far an observation is from the mean. This method can be used to identify outliers using a cutoff value. Z-score value calculated as:

$$Z - score = \frac{x - \mu}{\sigma} \tag{3}$$

Where, σ denotes the standard deviation of the sample, μ is the sample mean, and x is the data point.

However, removing outliers can influence study results, especially in a medical study.

Missing data is a common obstacle in the healthcare domain that can occur due to incorrect measurements, data entry errors, non-responses in surveys, and many more. Identifying the reason for the absence of data helps to handle it. Missing data can be grouped into three categories based on its form of missingness: Missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). These missing value categories determine the way to handle missing data. For example, data that are MCAR can be filled with a random sub-sample of actual data [63].

In general, missing data can be handled using imputation or deletion. It is safe to delete the data with MCAR or MAR without biasing the model. Nevertheless, removing data in NMAR can produce bias. There are three deletion methods for missing values.

1. List-wise deletion - Remove all observations that contain one or more missing values.
2. Pair-wise deletion - Remove variables with missing values and keep other variables with values.
3. Dropping variables - Remove variables if they have more than 60% missing data.

Imputation is useful when the percentage of missing data is low. This method uses substitute values to replace missing data. Missing value treatment methods depend on the variable type, whether it is categorical or continuous [64].

1. Categorical variables imputation

   Replacing missing data of a variable with the most frequent value (mode) is one of the basic imputation methods for categorical variables. Although mode imputation is an easy and fast method, it will introduce bias into the data set.

   Another simple method is to treat missing values as a separate category. Imputation can also be performed using predictive models, such as logistic regression, a random forest classifier, or a K-nearest neighbor (KNN), to estimate substitute values.

2. Continuous variable imputation

   The simplest continuous-variable imputation method uses the mean, median, or mode to impute the missing data. Although mean imputation is widely used, it ignores the correlations between variables. However, it is unbiased if MCAR and otherwise highly biased. Similarly to categorical variable imputation, it is possible to use machine learning models, such as regression models to predict missing values.

KNN model can handle both continuous and categorical data and is widely used to estimate missing values. The KNN method selects the k number of nearest neighbors to the missing data point and replaces the missing value with their average. To find the nearest neighbors, it uses different distance metrics based on the datatype; Euclidean, Manhattan, and Cosine distances for continuous data and Hamming distances for categorical data [65]. However, this method may provide low accuracy on high-dimensional data.

The hot and cold deck are the other two common imputation methods. Hot-deck imputation uses existing data to fill in missing values, where it replaces missing values with random values from that variable. Cold-deck imputation is similar to hot-deck imputation, except that external information or prior knowledge is used to replace missing data [63].

Inappropriate missing values handling may introduce bias, limit model generalizability, and lead to deceptive outcomes. Most ML models do not tolerate missing values; hence, proper missing value handling at the pre-processing stage is crucial. However, few models, such as Naïve Bayes and XGBoost, can deal with missing values internally.

Data transformation: After the pre-processing step, the data needs to be converted to an appropriate format. It involves data normalization, discretization, and generalization. Following these steps helps to improve the quality of the data. Several techniques are used to normalize data: min-max normalization, Z-score normalization, and decimal scaling. The normalization process scales the data into any arbitrary range, [a, b], for example ([0, 1] or [-1, 1]). This prevents variables with large ranges from outweighing others by giving equal weights to each variable. Some models may require data normalization, which helps accelerate the model training.

Min-max normalization scales the data to a decimal between the range a and b. It uses linear transformation on the original range and converts a value v of attribute V to a v' using,

$$v' = \frac{(v - V_{min})(b - a)}{V_{max} - V_{min}} + a \tag{4}$$

Where V_{max} and V_{min} are the maximum and minimum values of V, respectively [66]. While this method maintains the relationship among the original values, it is possible to create errors if future inputs are not in the initial data range. Another downside is that this method cannot handle outliers properly. However,

Z-score normalization can avoid this problem by transforming the values based on the mean and standard deviation of the feature using Equation 3.

The decimal scaling normalizes the data into the [-1, 1] range by moving the decimal point of the value of attributes. The value v_i normalized to v_i' by using:

$$v_i' = \frac{v_i}{10^j} \tag{5}$$

where j is the smallest integer such that $max(|v_i'|) < 1$.

Data analysis can be uncomplicated by ensuring the uniform scale on variables and simplifying information content using data discretization and data generalization techniques. During data discretization step, continuous data is converted into a set of data intervals with minimal data loss. For example, age can be discretized into three groups, 0-18, 18-55, and 55+. This method simplifies data representation and reduces memory usage; however, this process involves information loss. Data generalization, also known as blurring, is a process that converts low-level values into high-level concepts. For example, use the name of the city instead of the actual address.

Data reduction: This process reduces the original data volume using several techniques, such as dimensionality reduction, numerosity reduction, data cube aggregation and sampling. Reducing data improves the efficiency of the model, while producing the same analytical results.

The dimensionality reduction techniques aim to reduce the number of attributes in the dataset by eliminating redundant and irrelevant attributes. This can be done manually after recognizing the unimportant attributes. However, it becomes difficult for someone to figure out if the dataset is large. Hense, dimentionality reduction techniques are used to solve this problem and handle big data. Principal component analysis (PCA) is one of the popular linear-dimensionality reduction methods. It is an unsupervised method where it ignores the class labels. This method extracts important information from the data and expresses it as new orthogonal variables called principal components [67]. In contrast to PCA, linear discriminant analysis (LDA) is a supervised technique that tries to find the feature subspace that maximizes separability between classes [68].

Numerosity reduction brings down the original data volume by representing it in a smaller form. This technique is divided into two methods, parametric and non-parametric. Parametric methods store only model parameters, such as regression coefficients. Non-parametric methods use clustering, sampling, or histogram methods to reduce the representation of the original data. Data cube aggregation is another data reduction technique that aggregates multidimensional data into a simple form. Sampling reduces the number of observations in the dataset and this technique use to make larger dataset smaller. These methods accelerate the model training step and reduce the data storage space.

Data splitting: The final step of data processing is to split the pre-processed data set into training, validation and test data sets. The training dataset is used to train

the model, the validation dataset is used during the training phase to validate the model performance. It also helps to tune the model hyperparameters. Finally, the test data set is used to measure the performance and generalizability of the trained model. The test data set is independent from the training data. Typically, the 80:10:10% or 70:15:15% train-validation-test ratios are used to split the data, and in the ideal scenario the distributional properties will be the same in all three datasets.

3.2.2. Classification Algorithms

The next phase of the machine learning process is to train the learning algorithm using training data. ML algorithms have training parameters that are randomly initialized at the beginning. The target of the training is to find the best parameters of the algorithms that minimize the loss function, where the loss function computes the difference between the current output of the algorithm and the expected output. The model training of a supervised learning algorithm forms a mathematical representation of the relationship between features and class labels [69].

Currently, there are many machine learning algorithms, and it is crucial to select algorithms that are compatible with the problem at hand. Furthermore, algorithm complexity, efficiency, performance, and interpretability are additional factors to be considered. Several classification algorithms that are used for this thesis work are present in this section.

Logistic regression: This is a statistical method used to analyze data sets and to find the relationship between a dependent variable and independent variables. This algorithm is used as a classification method to predict a binary outcome based on historical data. Logistic regression can be further extended into three types based on the state of the target variable. Binomial; two possible classes, multinomial; three or more possible classes, and ordinal; three or more classes with a predetermined order.

The sigmoid function is the logistic expression used to map the predicted values to probabilities in logistic regression. This function converts any value into a value between 0 and 1 using the following formula:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

Where, $S(x)$ represents the probability, and x represents the predicted output.

This is a simple algorithm that requires much less training compared to other machine learning algorithms. Also, it is not affected by a small amount of noise in data. However, it is prone to overfitting, does not support nonlinear problems, and does not perform well if all independent variables are not identified [70].

KNN: This algorithm can be used to solve regression and classification problems. It is also used to impute missing values in the pre-processing steps. This method uses K nearest neighboring data points to predict the class of a new data point

[65]. It is a non-parametric model that does not assume any underline data distribution. KNN uses different distance measurements to find similar data points, as discussed in the data cleaning section.

Support vector machine (SVM): This algorithm is also used for regression and classification problems. During the training phase, hyperplanes are defined to separate data into classes. The points closer to this hyperplane from both classes are known as support vectors. SVM finds the best hyperplane that makes this decision boundary; the distance between the hyperplane and the support vectors is as wide as possible. SVM can handle linear and nonlinear problems; it can scale up with high-dimensional data and is less prone to overfitting. This algorithm has some downsides, as it does not work well with noisy data, does not provide prediction probability estimations, and performance can degrade with a large dataset [70].

Decision tree: This algorithm solves problems by continuously splitting data based on simple decision rules inferred from features. A decision tree is a graph that represents the decision rules on the nodes, the outcome of the decision on the branch, and the class labels on the leaf.

Figure 8 shows an example binary decision tree that predicts whether a person is fit or not fit based on the person's age and eating habits.

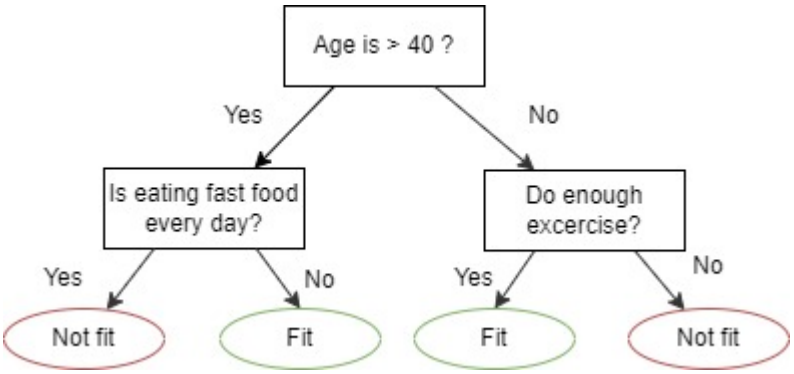

Figure 8. Binary decision tree: predicts whether person is fit or not.

The decision tree algorithm uses entropy to decide how to split the data [71]. Entropy is an information theory metric that measures the uncertainty or randomness in data, and it increases/decreases with the increase/decrease in uncertainty. The entropy value varies between 0 and 1. In a data set with a C number of classes, the entropy is calculated as;

$$Entropy(D) = \Sigma_{i=1}^{c} - p_i \log_2(p_i) \tag{7}$$

Where p_i is the probability of randomly selecting an example from class C, estimated by:

$$\frac{|C_{i,D}|}{|D|} \tag{8}$$

Information gain is another metric used in the decision tree training process. It measures the amount of information a feature provides about a class, and it helps to select the order of attributes in the decision tree. Information gain is the difference between original information gain and new requirements [71] calculated as:

$$Gain(D, A) = Entropy(D) - \Sigma_{j=1}^{v} \frac{|D_j|}{|D|} Entropy(D_j) \tag{9}$$

Where D denotes data partition, A denotes attribute, and v denotes distinct values in attribute A.

An attribute with the highest information gain is selected for the root node. This algorithm requires less data preparation, and it is easy to interpret and visualize. However, this algorithm is unstable and may involve expensive training due to complexity and higher training time.

Random Forest: This algorithm builds multiple individual decision trees and merges them to obtain more accurate results. It uses the bagging method to combine these multiple decision trees, creating different training subsets with replacements and using a majority vote for the final result [72]. Even though it is slower than the decision tree algorithm, it can efficiently handle large datasets and usually produces good results without hyperparameter tuning.

Naïve Bayes (Gaussian): This algorithm is a probabilistic classifier that uses the Bayes theorem. The theorem expresses the probability of the event based on prior knowledge of conditions that could be related to the event [73]. For a given predictor, the posterior probability of class c, $P(c|x)$, calculated as:

$$P(c|x) = P(x|c)\frac{P(c)}{P(x)} \tag{10}$$

Where $P(c)$ denotes the prior probability of a class, $P(x)$ denotes the prior probability of predictor, and $P(x|c)$ denotes the probability of predictor for a given class.

Naïve Bayes assumes that the predictors are independent of each other. Several algorithms are based on Naïve Bayes, such as Gaussian, Bernoulli, and Multinomial. Gaussian Naïve Bayes algorithm supports continuous values and assumes that all input variables follow the Gaussian distribution. The algorithm is highly scalable, not sensitive to irrelevant features, and able to work with less training data. The downsides of this algorithm are that data does not hold conditional independence assumption, and a zero probability problem could occur when the test data has specific class data that is not present in the training data.

Linear discriminant analysis (LDA): As discussed in the pre-processing section, this method is used for the dimensionality reduction method and as a supervised

classification technique. LDA projects a higher-dimensional feature space into a lower-dimensional space with a class of separable features. It helps to mitigate the curse of dimensionality by following three steps. It first calculates the between-class variance, the distance between the mean of different classes [74] using:

$$S_b = \Sigma_{i=1}^{c} N_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T \tag{11}$$

Where, c is the number of classes, N_i is the number of data points in each class, $\bar{X}_i$ is the mean of the i^{th} class, and $\bar{X}$ is the mean of the entire dataset.

Then, the distance between the mean and the samples of each class is calculated, called the within-class variance.

$$S_w = \Sigma_{i=1}^{c}(N_i - 1)S_i = \Sigma_{i=1}^{c}\Sigma_{j=1}^{N_i}(x_{i,j} - \bar{x_i})(x_{i,j} - \bar{x_i})^T \tag{12}$$

Where $x_{i,j}$ is the i^{th} sample in j^{th} class.

The final step, called Fisher's criterion used to form a transformation matrix P using the between-class variance and within-class variance. Calculated as follows;

$$P_{lda} = \arg\max_{P} \frac{|P^T S_b P|}{|P^T S_w P|} \tag{13}$$

Where P is the transformation matrix that is a projection of lower-dimensional space. Equation 13 can be written using eigenvalues (λ) of the transformation matrix P:

$$S_w P = \lambda S_b P \tag{14}$$

Equation 14 can be solved by calculating the eigenvalues (λ) and eigenvectors (v) of P if S_w is nonsingular;

$$P = S_w^{-1} S_b \tag{15}$$

Then calculate and sort the eigenvectors of P in descending order with their corresponding eigenvalues. First, k eigenvectors are then used as the lower-dimensional space.

LDA model estimates statistical properties for each class and uses the Bayes theorem to estimate the prediction probabilities. It assumes that each data point has the same variance and that the data have a gaussian distribution. LDA classification technique is used in many applications, including medicine, agriculture, and biometrics [74].

AdaBoost: Boosting is an ensemble modeling technique that combines a set of weak classifiers to obtain a strong classifier [75]. AdaBoost is a boosting algorithm,

and it is also known as adaptive boosting and most commonly uses decision trees or neural networks as component classifiers. This algorithm builds a model and initially gives equal weights to all data points. Then assigns higher weights to miss classified data points in the first model; these miss classified points will focus more on the next new model. The model creation and training will continue until receiving a lower error. These individual weak classifiers need to be aggregated to get a single strong classifier, commonly using a weighted average of the individual models. Adaboost performs with good generalization performance and is less vulnerable to overfitting. One of the main disadvantages of this method is that it requires quality data to provide good results.

XGBoost: This is a scalable algorithm for tree boosting that has recently been dominating in the ML research field [76]. XGBoost stands for extreme gradient boosting, and it provides a parallel tree boosting that helps to solve problems accurately and efficiently [77]. This is an extension of the gradient boosting method, which is another boosting technique. It uses the loss function to minimize the overall prediction error, regularization to smooth the final weights to avoid overfitting, shrinkage, and subsampling columns to prevent further overfitting.

This has proven to be one of the fastest algorithms compared to other algorithms due to its parallel and distributed computing. However, this algorithm is not suitable for small training datasets.

LightGBM (LGBM): Light gradient boosted machine is also an extended method of gradient boosting algorithm that uses decision tree models [78]. This algorithm grows a tree vertically (leaf-wise) by choosing a leaf with a large loss to grow, while other algorithms grow horizontally (level-wise). This helps to reduce the loss of leaf-wise algorithms. LightGBM is faster than the XGBoost algorithm, takes less memory, and is able to deal with large datasets.

It is crucial to select a relevant model for the problem at hand. After selecting a model, the model training step starts. During this step, the prepared data will be passed to the machine learning model to identify patterns in the data set and make predictions. The validation data set can be used to find the optimal hyperparameters that increase the accuracy of the model. After the training phase, model evaluation needs to be carried out with previously unseen data, the test data set.

3.2.3. Performance Evaluation Techniques

Model evaluation is a significant step in the ML pipeline. Performance metrics evaluate the model's accuracy and measure its performance. These techniques help to understand how well the trained model generalizes on the unseen data. Many evaluation metrics are available, and it is essential to choose suitable matrics. This section provides a theoretical background on the evaluation methods used in this thesis work.

Confusion matrix: Confusion matrix is an NxN matrix that summarizes the prediction results of a classification problem. It is not an evaluation metric, but most performance metrics are based on it. The confusion matrix is composed as shown in Table 1.

Table 1. Confusion Matrix

		Predicted values	
		Positive	Negative
Actual values	Positive	TP	FN
	Negative	FP	TN

Where,
TP - True Positive - correctly predicted positive class outcome
TN - True Negative - correctly predicted negative class outcome
FP - False Positive - incorrectly predicted positive class outcome
FN - False Negative - incorrectly predicted negative class outcome

Accuracy: Classification accuracy is a commonly used evaluation method that presents the percentage of correctly identified predictions.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{16}$$

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)} \tag{17}$$

Equation 17 demonstrate accuracy derived from the confusion matrix.

Accuracy can be used to distinguish a strong binary classifier from a weak one. However, it is not a proper method for imbalanced data because it does not identify correctly classified samples of different classes. Therefore, it may lead to incorrect results [79]. For this scenario, the balanced accuracy is used to evaluate the model.

Recall: Also called sensitivity. Measures the proportion of actual positive cases that are correctly classified.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

Precision: Precision is another metric that can be used for an imbalanced data set. It calculates the proportion of positive cases that are identified correctly.

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

Specificity: Also called as true negative rate (TNR). This method calculates the correct negative predictions, where the best specificity is 1, and the worst is 0.

$$Specificity = \frac{TN}{FP + TN} \tag{20}$$

Balanced Accuracy: Balanced accuracy is a more suitable metric for model evaluation when dealing with an imbalanced data set. It deals with both positive and negative outcomes in binary classification and is calculated as follows:

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{21}$$

F1-score: This metric is useful when the class distribution is uneven. It keeps the balance between precision and recall by calculating their harmonic mean as:

$$F1 - score = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right) \tag{22}$$

F1 score gives a high/low value if both the precision and recall values are high/low.

Area under the ROC (AUROC) curve: ROC stands for receiver operator characteristics, and the ROC curve is a plot of recall (sensitivity) against the specificity that shows the trade-off between them.

The AUROC curve measures the ability of a model to distinguish between positive and negative classes. A higher value of the area under the curve (AUC) implies better model performance. This method can use for balanced data as well as imbalance datasets.

Threshold moving for class imbalance classification: Many machine learning models predict observation's probability of belonging to available classes and convert probabilities to the class label. This method provides a way to measure prediction uncertainty. This mapping is achieved using a threshold of 0.5 in binary classification. All the observations with a probability equal to or greater than 0.5 will assign to one class, and all the others are mapped to another class. When a classification problem is carried out with a class imbalance dataset, such as a medical dataset, this default 0.5 threshold can provide poor results. Threshold tuning can be used to improve the performance of a classifier that predicts the probabilities of a class imbalanced dataset. The optimal threshold can be calculated using ROC curves, Precision-Recall curves, or manually.

The ROC curve can explain the trade-off in true positive rate (TPR) and false positive rate (FPR) for different thresholds. The objective is to locate the threshold with the optimal balance between TPR and FPR [80]. The geometric mean is a metric that is used to find the balance between TPR and FPR, calculated as:

$$G - Mean = \sqrt{(TPR * (1 - FPR))} \tag{23}$$

Select the threshold with the largest G-mean value as the optimal threshold.

The precision-recall curve concentrates on classifier performance in the minority class. It uses the F-score to find the optimal threshold. Select the threshold with the largest F-score [80].

3.3. SHAP

Healthcare is a sensitive research area that deals with human lives. All decisions must be taken with care and with solid evidence. The availability of big data enables the facility to use more complex models that provide more accurate results. However, this comes with the weakness of lack of transparency. In the medical field, the output of machine learning solutions is not accountable if it cannot be adequately explained. It is crucial to interpret the model output and understand the reasons behind the model's decision. One of the main reasons for neglecting some ML models in the real world is the lack of interpretability of models [81]. Various methods [82, 83] have been proposed to interpret the predictions of the complex model, and in 2017 Lundberg and Lee [84] addressed this problem by presenting a unified framework named SHAP, which interprets the predictions of a model. SHAP is a Python library, and it stands for SHapley Additive exPlanations.

This library uses the Shapley value, a cooperative game theory concept, to achieve transparency and interpretability of the model. In the game theory, the Shapley value is the average marginal contributions across all permutations [85]. Shapley values consider the prediction task for a single instance as the 'game', and the feature values of the instance as 'players' to interpret the machine learning model. SHAP calculates the Shapley values for each feature of the sample, where it represents the feature impact in the generated prediction.

SHAP models are built on the training dataset. `shap.Explainer` is the primary interface in the SHAP library that explains any machine learning model. There are several other Explainers implemented from `shap.Explainer` interface, such as `TreeExpainer`, `DeepExplainer`, and `KernelExplainer`. These explainers support different ML model types. TreeExplainer uses a tree SHAP algorithm to explain tree-based machine learning models and ensembles of tree-based models. XGBoost, LightGBM and CatBoost are some models that can be explained using TreeExplainer. DeepExplainer explains deep learning models, and KernelExplainer can be applied to any model. These explainers may provide different results since the underline algorithms are different.

SHAP provides global and local interpretation methods to explain the output of the model. All the SHAP plot illustrations in this section are based on the Iris data set [86]. This dataset contains 150 samples with four features (length and width of petals, length, and width of sepals) of three Iris species (Iris-setosa denoted by class 0, Iris-versicolor by class 1, and Iris-virginica by class 2). An SVM model is built to classify the Iris species in this dataset. SHAP plots are used to interpret this model's predictions on the Iris dataset in the following sections.

3.3.1. Global Interpretability

Global interpretation methods show how much each feature contributes to the model prediction, either negatively or positively. Several global interpretation methods/plots are available in SHAP, which are based on aggregations of Shapley values.

Feature importance: This method plots the most significant features in decreasing order using average absolute Shapley values per feature. Features on top of the plot contribute more to the model prediction, and the bottom ones have less predictive power.

Figure 9 illustrates the feature importance plot where petal length has the most substantial influence on prediction in all three classes. Sepal length and sepal width have a low influence on prediction.

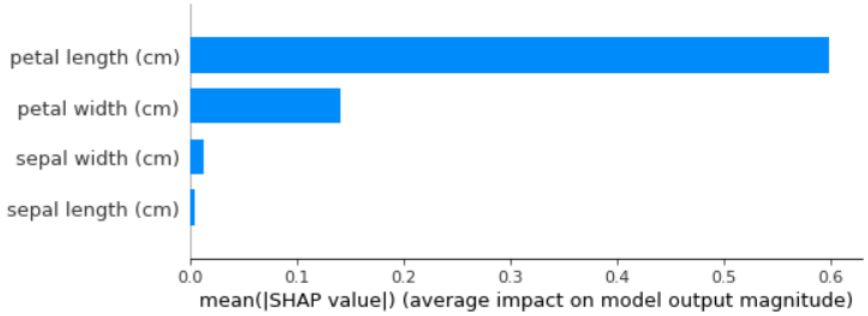

Figure 9. SHAP summary plot: Feature importance.

Summary plot: This visualizes an information-dense summary of how features impact the model output. It is a combination of feature importance and feature effects. Similar to the feature importance plot, features are ranked based on their importance. The X-axis represents the Shapley values, and the Y-axis represents the features.

Each sample is represented by a single point on each feature row. Its x-position represents the Shapley value of that feature, and the point color represents the original value of the feature, with red for high values and blue for low values. Some models allow missing values, and in that scenario, the missing values in the dataset will represent with gray color. The horizontal location of the point shows the effect of that value on prediction, where points with positive Shapley values provide a higher prediction effect, and negative points show a low prediction effect.

Figure 10 illustrates the density scatter plot of the SHAP values for the Setosa output. Petal length is the most important feature on average, and the least important feature is sepal length. The petals of low length have a positive impact on increasing the probability that the sample is classified as Setosa with high SHAP values. This low length is derived from the color of the data points and positive impact on Setosa classification because data points are on the positive side of the X-axis. Similarly, low petal width and sepal length positively impact Setosa classification, while sepal width is negatively correlated.

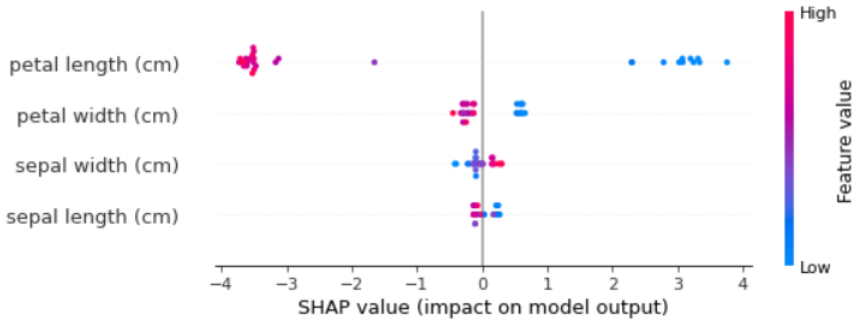

Figure 10. SHAP summary plot: Information-dense.

Dependence plot: This scatter plot shows how a single feature effect model prediction. Visualize each data point with the feature value on the x-axis and the corresponding Shapley value on the y-axis. The color of each point represents the original value of the second feature that may have an interaction effect with the current feature. It is possible to set a feature for this second feature; otherwise, the tool will automatically select one.

Figure 11 depicts the Setosa output showing a negative relationship between the petal length and the target variable. Low petal lengths show a higher probability of being Setosa. Petal length and petal width frequently interact, where the color shows the original values of the petal width.

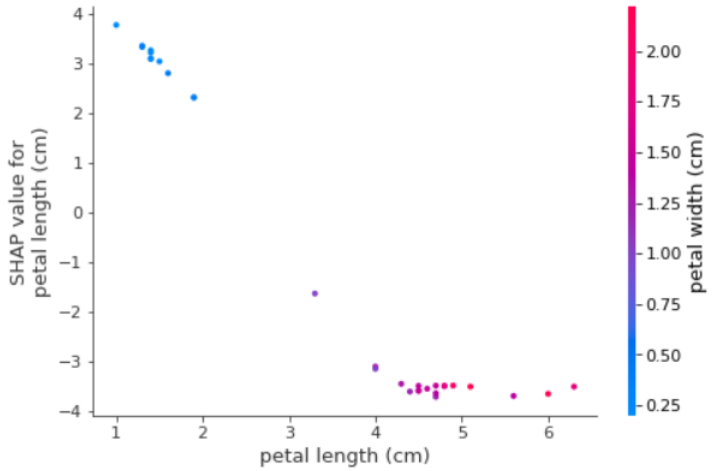

Figure 11. SHAP dependence plot.

3.3.2. Local Interpretability

Traditional methods only provide global interpretation methods, such as feature importance. These methods show results across the entire dataset and not on each observation. SHAP introduced method, $force\_plot$, that summarize how each feature contributes to predicting each observation in the dataset. This method highly increases the transparency of model prediction, especially for predictive models in the medical field to explain the reasons behind the predicted output for a particular patient [85].

Each observation has its own set of SHAP values for corresponding features and the x-axis on the this plot is in log-odds space. This plot uses those SHAP values to explain why that observation receives its prediction and feature contribution.

Figure 12 explains the Setosa output prediction of the first sample in the test dataset, where the actual class of the first data sample is Virginica. The scale represents the model prediction values and the output value, 0.01, is the prediction for the sample to be Setosa. The base value indicates the mean prediction. The features that push the prediction to the left (lower) are colored blue, and those that push it higher are colored red. The blue indicates that the feature decreases the probability of classifying this data sample as Setosa. In the figure, all the four features negatively impact classifying this sample as Setosa, where petal length is 5.1cm.

This SHAP local interpretation plot may not be clear with the presence of high number of features. Figure 13 illustrates a more readable and improved version of the original SHAP $force\_plot$ (Figure 12). This bar plot is not available in the SHAP library and it was developed for this study using the SHAP values. It exhibits the same interpretation as $force\_plot$. However, the scale in Figure 12 represents the model prediction, while the X-axis in Figure 13 represents the SHAP values.

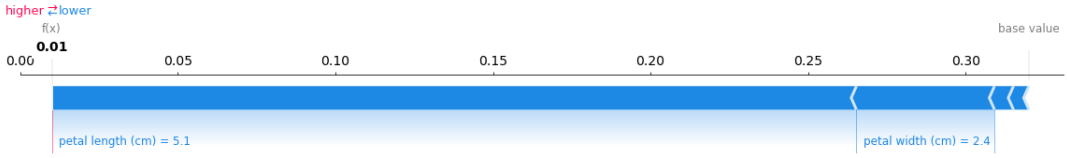

Figure 12. Interpretation of Setosa prediction for the First sample on test dataset.

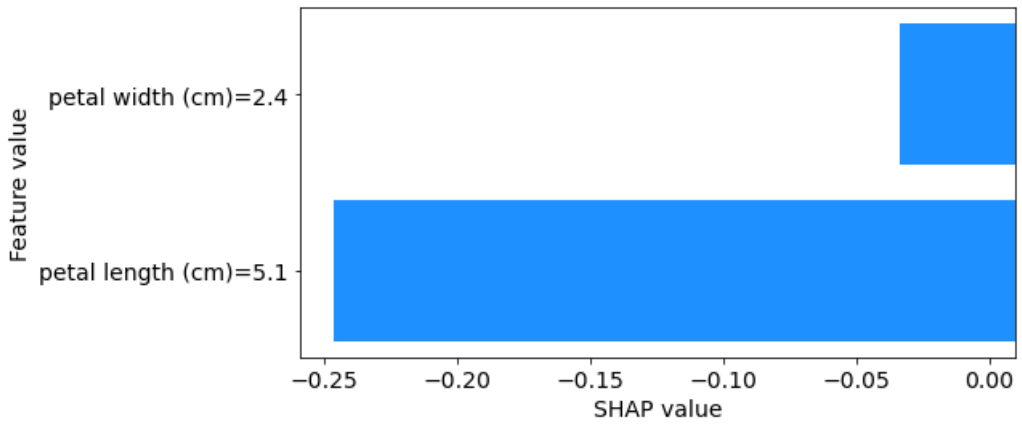

Figure 13. Detailed illustration of Figure 12.

Figure 14 illustrates the Virginica output prediction for the same sample. All the four features show a positive impact to predict this sample as Virginica, where features push the prediction to the right side with a 0.96 output value. A more detailed illustration of Figure 14 can be seen in Figure 15.

SHAP provides several other plots that help to interpret model outcomes, such as an image plot that provides SHAP values for image input and a text plot that explains

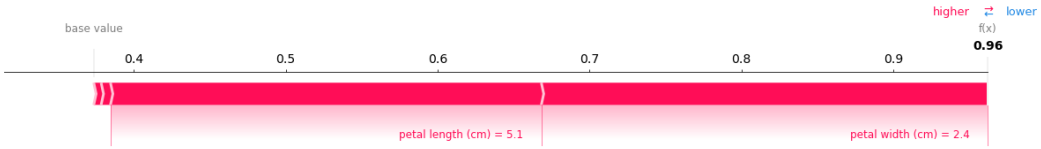

Figure 14. Interpretation of Virginica prediction for the First sample on test dataset.

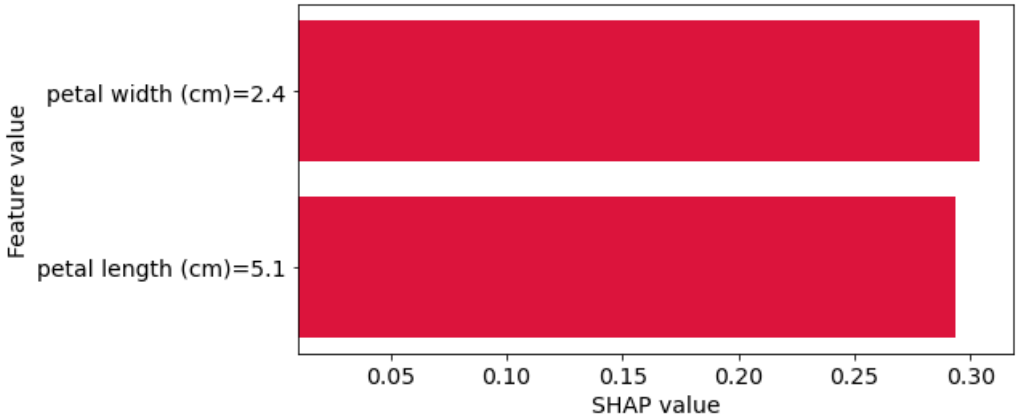

Figure 15. Detailed illustration of Figure 14.

a text using labels and colors. SHAP is a powerful tool that helps to understand the behavior of the machine learning model.

4. EXPERIMENTS AND RESULTS

This chapter provides a detailed description of the prediction models that identify the possible risks of SH and DKA events in T1D patients with 12 months prediction horizon. In the initial step, the models were developed using the features mentioned in the replication study. These initial models were further improved by performing several experiments. In addition, the system interprets the prediction models and their individual outcomes. This interpretation helps to understand the model decision-making process better. Finally, prediction models were used to build a decision support system to identify patients at high risk for developing SH and DKA events. All of the work in this thesis was carried out in Python.

4.1. T1D Exchange Data Pre-Processing

The enrollment data set in the T1D exchange registry was used for this thesis work. The data is stored in five separate text files: subjects, medical conditions, lab results, HbA1C test results, and medications. The subjects-file includes one record per patient and contains patient data, visit data, family history, quality of life, and other vital data. The HbA1c-file contains the patient's HbA1c data for the past 12 months.

The data files were loaded and converted to DataFrames using the Python pandas library. The first pre-processing step was carried out to integrate medical conditions, medications, and mean HbA1c data (calculated) with the subject data. All duplicate observations and insignificant features were removed from the integrated dataset.

The dataset contained a considerable amount of inconsistent data entries,

1. White space inconsistencies were present in categorical variables. It was fixed by removing the white spaces at the beginning and end of the labels.

2. Capitalization inconsistencies were removed by making every label to lowercase.

3. The dataset contained different categorical labels with the same meaning; for example, both *'DK'* and *'Don't know'* labels were used when patients did not know the answer. Similarly, *'Yes'*, *'1.Yes'*, and *'1.Y'* labels indicate *'Yes'*. Convert all these different labels into one consistent label.

4. Labels contain mathematical notations such as '>', '<', '='. For example, the $Pt\_NumHospDKA$ feature has numerical values except *'>9'*, and the $AgeAtConsent$ feature contains integer values except for the *'90 or older'* string. Replace all these inconsistent values with the smallest possible integer.

Machine learning models cannot interpret categorical features, and it is required to convert categorical data into the numeric format in the pre-processing step. All ordinal, nominal, and Boolean categorical features in the data set were converted into numeric values.

The data set has features with one possible value. These feature data were collected from the participant questionnaire forms, where a checkbox is provided to allow users to make a binary choice. For example, a feature named $Pt\_InsPriv$ denotes that the

patient has private health insurance or health care coverage. The only possible value in this feature is 1, which indicates that the person has private health care coverage. All patients who do not answer this question have missing values for this feature. The missing values in the dataset are defined as empty strings. Since these data were collected from participants, it is safe to assume that if a person uses private insurance, he/she will fill out the checkbox in the form correctly. Assuming that, filled missing values in the significant features that have only one possible value (value 1) with 0. In the next step, all insignificant features that contain more than 60% missing values were removed. Mode imputation (most frequent value) was used to replace missing values in other features.

A feature engineering step was carried out to derive new features from existing features. The enrollment data set does not contain the T1D duration data field. Hence, it was derived using the patient's age at diagnosis and age at consent. The data set contains features that can be used to derive if the person is going through a stressful situation. A new feature was introduced by combining features that represent the main stressors of life, such as stress related to studies, work, family, financial, legal, close ones' death, or marriage life. The patient's family history with T1D was used to create a new feature $relative\_T1D$. It was updated if the patient's immediate biological family members are present with T1D, including father, mother, child, siblings, twin, grandparent, or grandchildren.

Overweight and obesity are significant concerns in T1D patients that can be determined using the body mass index (BMI). Dataset does not include BMI data, and it was derived from the height and weight of the patients. The patient's height and weight measurements are available at enrollment and when they have been diagnosed with T1D. Weight measurements can be found in kilograms or pounds, and height measurements in inches or centimeters. These features are converted into one standard unit: kilogram for weight, centimeter for height, and used to calculate the body mass index (BMI) feature. After the feature engineering step, redundant and irrelevant features in the dataset were eliminated. Highly correlated features were also excluded from the dataset.

This work was carried out with the similar patient group mentioned in the replication study [54], patients aged 26 years or older and had a T1D duration of at least two years. The cohort consists of 7156 individual patients' data after filtering using patient age (>= 26) and T1D duration (>=2).

The target variable for predicting SH events ($SHSeizComaPast12mos$) describes whether the participant had a severe hypoglycemic event that occurred in the last 12 months. This feature has three possible values: *'Yes'*, *'No'*, or *'Unknown'*. All the *'Unknown'* observations in $SHSeizComaPast12mos$ were removed to make it a binary feature. $NumSHSeizComaPast12mos$ (Number of known severe hypoglycemic events in the last 12 months) feature knowledge was used to fill the possible missing values in the target feature. If $NumSHSeizComaPast12mos$ is greater than zero, the missing value was filled in with *'Yes'* considering that the patient had an SH event in previous year.

Similarly, all *'Unknown'* observations in the target feature for DKA prediction were removed to make it binary. The target feature, $DKAPast12mos$, describes whether the patient had DKA in the last 12 months. The $NumDKAOccur$ feature (number of

known DKA occurrences in the last 12 months) was used to fill in the possible missing values in the target variable.

The features $SHSeizComaPast12mos$ and $NumDKAOccur$ were removed from the dataset since these features influence the prediction. After the preprocessing step, the data set contained 205 individuals with at least one SH event and 6450 individuals with no SH events in the previous 12 months. Furthermore, 200 individuals had at least one DKA event in the previous 12 months and 5971 individuals had no DKA occurrences during the period. Additionally, 12 patients had both SH and DKA events in the previous 12 months.

The final preprocessing step was carried out to split the dataset into training and test sets with a 3:1 ratio. This preprocessed dataset was used to implement all the prediction models in this work.

4.2. Prediction Models with Prior Study Findings

This section describes the implementation details of SH and DKA prediction models using the same study group and factors mentioned in the study that we are replicating [54]. Both SH and DKA prediction models were built with the same set of features presented in Table 2. It includes features that showed association in both univariate and multivariate models in [54], including patients' socioeconomic status, such as annual income, private insurance, and education level.

Table 2. Feature description

Feature Name	Description	Unit / possible values
diagDuration	Duration of T1D	Years
HbA1c	Mean HbA1c in the past year	Percentage (%)
Gender	Gender	Male Female Transgender
Pt_RaceEth	Race/Ethnicity	Black/African-American Hispanic or Latino Native Hawaiian/Other Pacific Islander Asian American Indian/ Alaskan Native More than one race
Pt_AnnualInc	Annual household income from all sources	< $25,000 $25,000 - < $35,000 $35,000 - < $50,000 $50,000 - < $75,000 $75,000 - < $100,000 $100,000 - < $200,000 $200,000

Pt_InsPriv	Private health insurance	1 or 0 (Yes/No)
Pt_EduLevel	Highest level of education participant completed or highest degree achieved	< Less than 1st Grade 1st, 2nd, 3rd, or 4th grade 5th or 6th grade 7th or 8th Grade 9th Grade 10th Grade 11th Grade 12th Grade - no diploma High school graduate/ diploma/ GED Some college but no degree Associate Degree Bachelor's Degree Master's Degree Professional Degree Doctorate Degree
BMI	Body mass index	Float value
Pt_InsulinRecMethod	Insulin received method	Pump Injections/pens Pump and injections/pens Sometimes pump and sometimes inj/pens Do not take insulin
Pt_NumBolusDay	How many boluses per day of insulin	Interger (0-35)
relative_T1D	First degree relative with T1D	1 or 0 (Yes/No)
Pt_SmokeAmt	Current smoker	1 or 0 (Yes/No)

A correlation matrix can explain how the variables are related to each other. It depicts the correlation between all the possible feature pairs. Correlation coefficients greater than zero show a positive relationship between features, and a value less than zero indicates that there is a negative relationship between features. Figure 16 shows the 'lower' triangle of the correlation matrix.

Private insurance, level of education, and relatives with T1D features positively correlate with the SH target feature. Ethnicity, annual income, daily insulin units, insulin delivery method, and relatives with T1D features negatively correlate with the DKA target variable. A high correlation can be seen between education level and private insurance features. Age, education level, and private insurance features are positively correlate with the current smoking state variable.

The sub-chapters in this section describe the experiments and results carried out in this work. Since the data set is imbalanced, all the experiments used the threshold moving technique to achieve the optimal threshold that separates the two classes. Ten different models were trained for each case. These ten models are LightGBM,

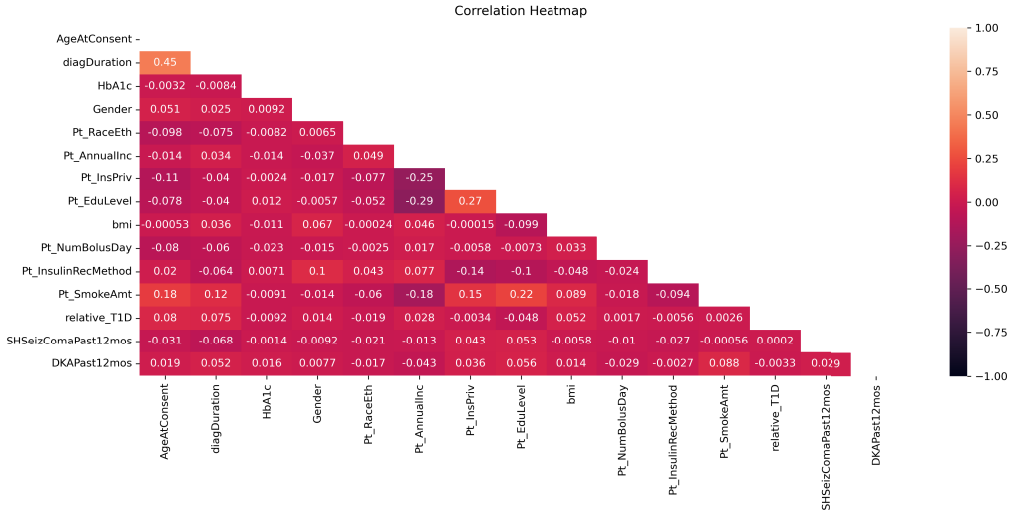

Figure 16. Correlation heatmap: correlation between selected variables.

XGBoost, AdaBoost, Random Forest, Decision Tree, Logistic regression, Linear discriminant analysis, Gaussian Naïve bayes, SVM and K-nearest neighbour classifier. Both LightGBM and XGBoost classifiers allow to have missing values. These two classifiers were used with and without missing value imputation methods. In this chapter, only the two best results will be reported for each case. Cross-validation was used to assess the model's effectiveness. Class-wise accuracy, balanced accuracy, Area under ROC, and F1-score (micro) were used as evaluation parameters to provide a more robust comparison.

4.2.1. SH Prediction Models

The models were built to predict possible SH occurrences in adult patients with T1D within the next 12 months. The objective is to use factors that showed a significant association with SH events in a replication study [54] to predict possible SH events. Ten different models were trained with predictors mentioned in Table 2 and used 10-fold cross-validation to evaluate the estimator performance. During the 10-fold cross-validation, the training data set was divided into ten subsets, and each of the subsets was used to evaluate the model fit in the other nine subsets. The performance of the selected models was tested using the test data set.

Correctly classifying individuals with SH occurrences (TP) is more important than identifying individuals without SH events (TN). From trained 10 models, the two best results are display in Table 3. Both the XGB and LightGBM classifiers trained with a dataset that contains missing values provided the best results with the baseline factors.

LightGBM classifier was able to identify individuals having SH events with 76% accuracy. Figure 17 illustrates the relative importance of each feature when the model makes a prediction. This feature importance's are defined using Python Scikit-learn library [87], that calculate scores for all the predictors in the model. The duration of T1D, the patient's age, and BMI significantly impact the model predictions.

Table 3. Results of SH prediction with baseline factors

Model	Class wise accuracy		Balanced accuracy		Predicted labels		F1 score - micro	ROC curve
	SH	Non SH			Positive (P)	Negative (N)		
LGBM	0.76	0.57	0.66	P	39	12	0.72	0.66
				N	692	921		
XGBoost	0.66	0.59	0.62	P	34	17	0.73	0.62
				N	657	956		

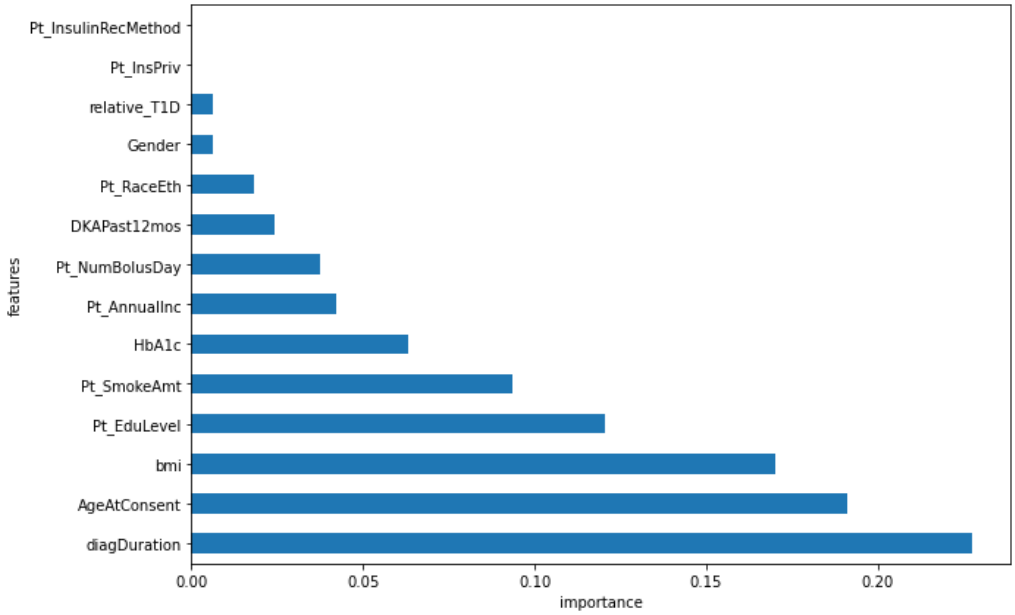

Figure 17. Feature importance: LGBM classifier - SH prediction with baseline factors.

4.2.2. DKA Prediction Models

The prediction model was built to predict possible DKA occurrences in the next 12 months. The same baseline factors were used for predictors and the models were trained with a ten-fold cross-validation. The two best results are displayed in Table 4.

The AdaBoost classification model achieved the best results in predicting possible DKA events with 67% balanced accuracy and with 70% accuracy for the DKA class. Both models achieved same F1 score. However, it is more important to identify positive cases of DKA, and balanced accuracy is a better metric in this case.

Figure 18 summarizes the feature importance of the AdaBoost model, where the duration of T1D, the mean HbA1C, age, and BMI express the impact on the prediction of DKA events.

Table 4. Results of DKA prediction with baseline factors

Model	Class wise accuracy		Balanced accuracy		Predicted labels		F1 scor - mcro	ROC curve
	DKA	Non DKA			Positive (P)	Negative (N)		
AdaBoost	0.70	0.64	0.67	P	35	15	0.64	0.67
				N	527	966		
LGBM	0.60	0.64	0.62	P	30	20	0.64	0.65
				N	529	964		

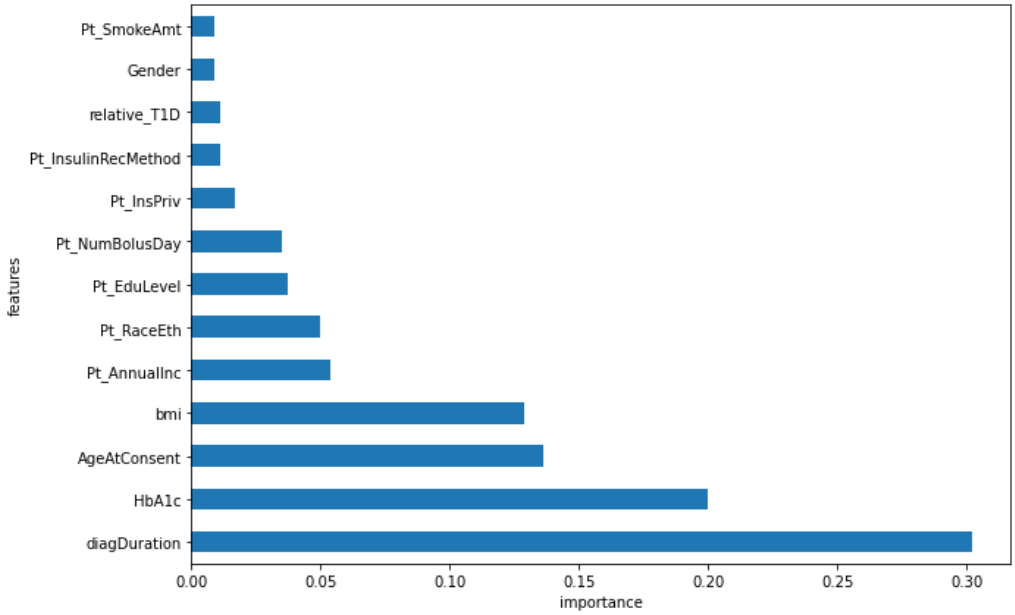

Figure 18. Feature importance: AdaBoost classifier - DKA prediction with baseline factors.

4.3. Prediction Models with Improvements

The prediction models presented in 4.2.1 and 4.2.2 used only the factors found by Weinstock et al. [54]. We used 13 features to forecast the T1D outcomes where both models showed poor performance. These associations alone are not sufficient to properly predict SH and DKA events. After carrying out the experiments below, we were able to improve the models further.

1. The preprocessed dataset contains 208 features that provide valuable information about the patient's condition. Experiments were carried out using all these features as predictors to build the models.

2. The feature selection method was used to select features from the preprocessed dataset and build models using selected features.

3. The dataset contains information that is only relevant to female patients, such as when they first had a menstrual period, whether they have a regular menstrual cycle or not, the reason for the irregular menstrual cycle, current pregnancy status, and the number of miscarriages.

   Built separate predictive models for Males and Females and removed features that are only applicable to females from the male dataset.

   The final dataset contained 3615 female patient records, 3038 male and two transgender records. 106 females reported SH event occurrences, and 123 females reported DKA occurrences in the past 12 months. 99 male patients reported SH events, and 77 reported DKA events. Transgender records were removed from the dataset since those two patients do not report any SH or DKA event occurrences.

This chapter covers the SH and DKA prediction models that produce the best results after carrying out the above experiments.

4.3.1. SH Prediction Models

The best results were achieved with gender-disaggregated prediction models. Medical conditions data collected from the questionnaire form were removed from the pre-processed data set except for anxiety, diabetic peripheral neuropathy, high LDL, low HDL, and osteoporosis conditions. The remaining 147 features were used to predict possible SH events.

Prediction model - male population: LightGBM model correctly classified all individuals with SH events in the test dataset and achieved 86% balanced accuracy (Table 5).

Table 5. Results of SH prediction - Male population

Model	**Class wise accuracy**		**Balanced accuracy**		**Predicted labels**		**F1 score - micro**	**ROC curve**
	SH	Non SH			Positive (P)	Negative (N)		
LGBM	1.0	0.72	0.86	P	20	0	0.73	0.84
				N	205	530		
XGBoost	0.80	0.83	0.81	P	20	5	0.83	0.84
				N	123	612		

Figure 19 illustrates the 30 most important features of the LightGBM model. The patient's history of previous SH events ($Pt\_NoSevHypoEver$ - never had SH events in his life) greatly influenced the prediction. Number of times patient visited a healthcare provider (physician or nurse practitioner) to

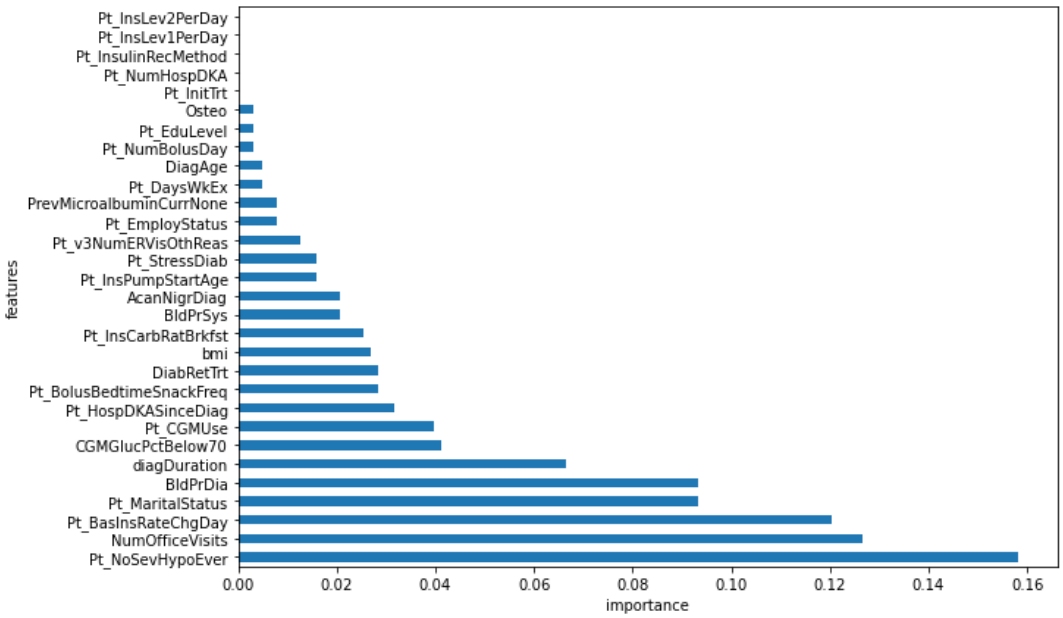

Figure 19. Feature importance: LGBM classifier - SH-Male model.

review diabetes management ($NumOfficeVisits$), number of times per day a patient basal insulin rate change ($Pt\_BasInsRateChgDay$), current marital status ($Pt\_MaritalStatus$), diastolic blood pressure ($BldPrDia$) and diabetes duration ($diagDuration$) shows significant impact on the prediction.

Prediction model - female population: Both LightGBM and XGBoost classifier produced similar results in identifying SH events in female patients (Table 6). However, the LightGBM classifier model showed a slight increase in the accuracy of identifying non-SH classes.

Table 6. Results of SH prediction - Female population

Model	**Class wise accuracy**		**Balanced accuracy**		**Predicted labels**		**F1 score - micro**	**ROC curve**
	SH	Non SH			Positive (P)	Negative (N)		
LGBM	0.85	0.72	0.78	P	23	4	0.72	0.80
				N	243	634		
XGBoost	0.85	0.71	0.78	P	23	4	0.72	0.82
				N	246	631		

Similar to the male prediction model, the history of SH events, the number of visits to the healthcare provider, the duration of diabetes, and the diastolic blood pressure are among the six most important features of the model. Furthermore, current pregnancy status ($Pt\_currpreg$), currently having regular menstrual or not ($Pt\_RegMenstCyc$), reason for irregular menstrual cycle

($Pt\_IrregMenstCycReas$), height, and weight influence the prediction in female population (Figure 20).

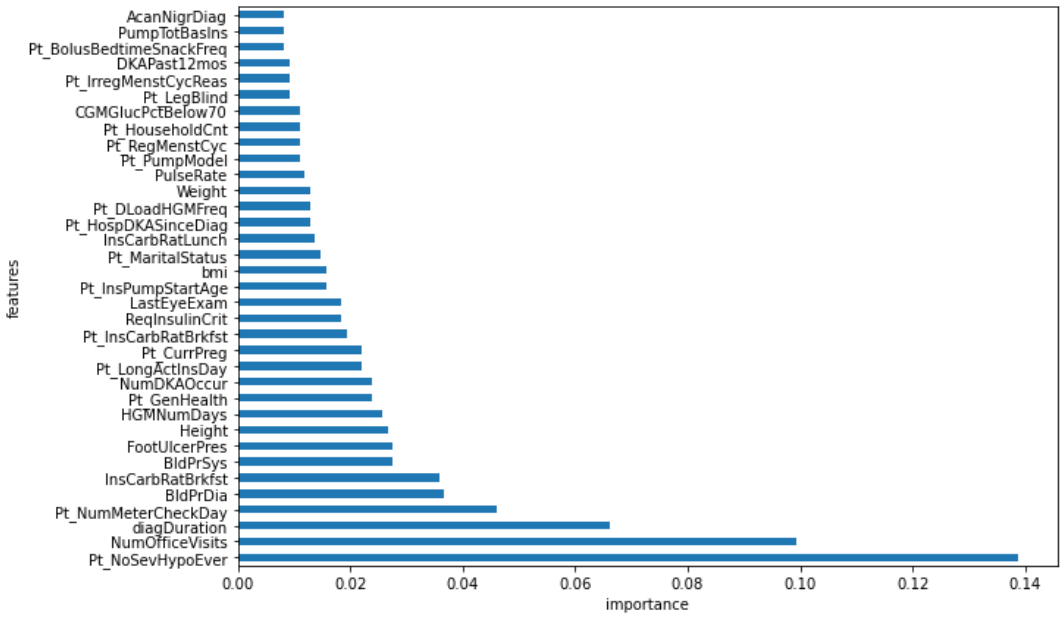

Figure 20. Feature importance: LGBM classifier - SH-Female model.

A considerable number of factors mentioned in [54] were prominent in both male and female SH prediction models. Diabetes duration and BMI impacted both models, while education level, employee status (related to income), and the number of insulin boluses per day were important features for the male SH prediction model.

4.3.2. DKA Prediction Models

The results obtained from a single model and gender-specific models were compared to identify the best model. A single prediction model that used all the features in the preprocessed data set achieved the best result for predicting DKA events. LightGBM classifier achieved 88% accuracy in predicting DKA occurrences with 82% balanced accuracy, while AdaBoost gave 80% balanced accuracy (Table 7).

Figure 21 illustrates the feature importance of LGBM classifier. The feature indicates that the patient's history of hospitalization overnight for DKA with high blood sugar and ketones ($Pt\_HospDKASinceDiag$) significantly impacts model prediction. The weight, duration of T1D and annual household income ($Pt\_AnnualInc$) also show a significant impact on the prediction of the model.

4.4. Model Interpretation

Medical field values model interpretability; the ability of the model to explain the reasons for receiving such a prediction. Real-world prediction models that deliver the best results are not trustworthy in the medical domain if they cannot explain the

Table 7. Results of DKA prediction

Model	Class wise accuracy		Balanced accuracy		Predicted labels		F1 score-micro	ROC curve
	DKA	Non DKA			Positive (P)	Negative (N)		
LGBM	0.88	0.77	0.82	P	44	6	0.77	0.87
				N	340	1153		
AdaBoost	0.80	0.81	0.80	P	40	10	0.81	0.84
				N	280	1213		

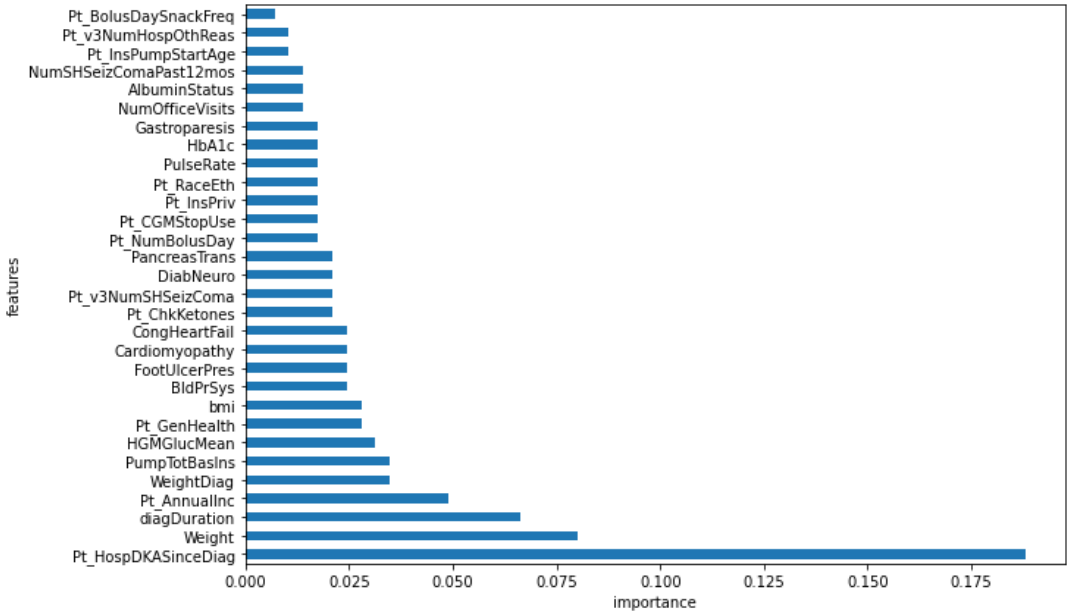

Figure 21. Feature importance: LGBM classifier - DKA model.

result. The built models classify SH/Non-SH and DKA/Non-DKA occurrences in T1D patients, and it is important for both healthcare provider and patient to know why the patient got classified with possible outcome risk. Knowing the factors will help them manage the diabetes. In this thesis, we are trying to bridge the interpretability gap in medical machine learning using the SHAP library. Global interpretation is achieved by using the SHAP information dense summary plot. That gives a view of how features and their values impact on model outcomes across the whole dataset. It is important to have a reason for each individual's outcome separately. Local interpretability is achieved by reviewing the Shapley values of each prediction. It expands the interpretability of the model by predicting possible risks and providing the reasons behind the prediction.

4.4.1. SH-Male Prediction Model Interpretation

LightGBM classification showed the best results in classifying males with SH occurrences. Imputation methods were not used since the LightGBM model handles missing values. This section provides a global and local interpretation of this model using the test dataset.

Global interpretation: Figure 22 illustrates how the top 14 feature values impact the model output. The LightGBM model allows to have missing values and these missing values are represented by gray points. Data samples with negative Shapley values show a higher prediction effect for SH occurrences, and points with positive Shapley values show a low prediction effect for SH events. Male patients with a previous history of severe hypoglycemia (blue points) are more likely to have SH occurrences. Furthermore, patients with fewer changes in the basal insulin rate (per day), a high number of visits to healthcare providers, diabetic retinopathy, low T1D duration, patients who are divorced or widowed, patients that are often stressed, are more likely to develop SH occurrence within the next 12 months.

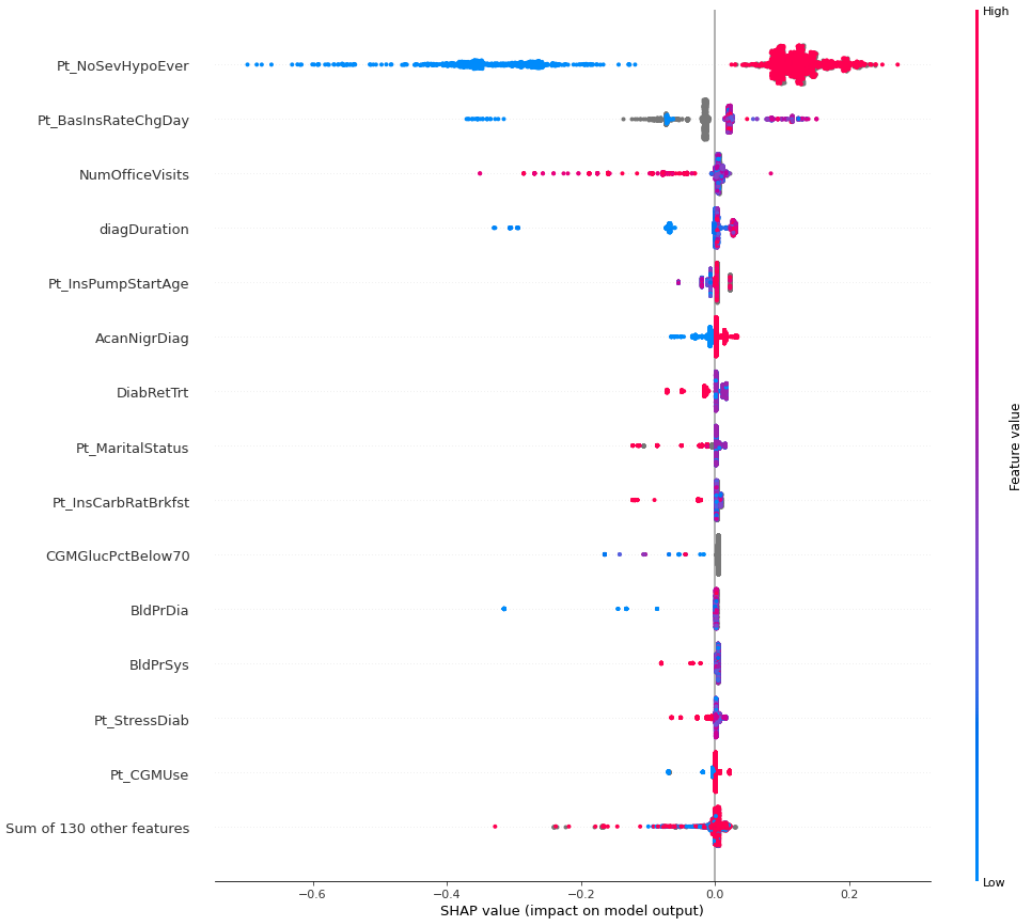

Figure 22. Global interpretation - SH prediction of male population.

Local interpretation: The local interpretation tries to explain the reasons for the individual prediction, whether the patient has an SH event or not, and how

the features contribute to making the decision. The output SHAP values with negative decisions (non-SH events) will be on the right side of the base value, and positive decisions will be on the left side. Figure 23 interprets the model decision for the true negative case in the test dataset, where the patient described by the sample does not have any experience with SH events, and the prediction model correctly identifies him as a sample without SH events. To improve the readability of Figure 23, we expanded the original SHAP plot. Figure 24 illustrates the expanded plot where it exhibit the same interpretation as Figure 23 with additional information. However, the X-axis in $force\_plot$ is in log-odd space and we converted log-odd values to probability in Figure 23, such that the scale represents the model predicted values that are converted into probability. Moreover, the X-axis in Figure 24 represents the SHAP values.

The model prediction illustrated in Figure 23 and Figure 24 belongs to a married person ($Pt\_MaritalStatus$ = 2) who does not have a history of severe hypoglycemia ($Pt\_NoSevHypoEver$ = 1), and currently uses a real-time continuous glucose monitor (Pt_CGMUse = 1). In a typical day, his basal insulin rate change 2 times ($Pt\_BasInsRateChgDay$ = 2) and most of the time he takes bedtime snacks ($BedtimeSnackFreq$ = 3). These are the main reasons for the model to decide that the patient will not be at risk of having an SH event in the next 12 months. The history of SH events and the use of continuous glucose monitoring significantly support the decision. The reasons indicated in blue, that he has visited the health provider seven times during the previous year ($NumOfficeVisits$ = 7), and the diabetic retinopathy condition mentioned as 'Unknown' ($DiabRetTrt$ = 2) increase the probability of classifying him as a positive case. However, the SHAP values for negative decisions (non-SH events) show a significant impact on this patient, where the predicted value (0.02) is greater than the base value (0.014), classifying him as a non-SH patient.

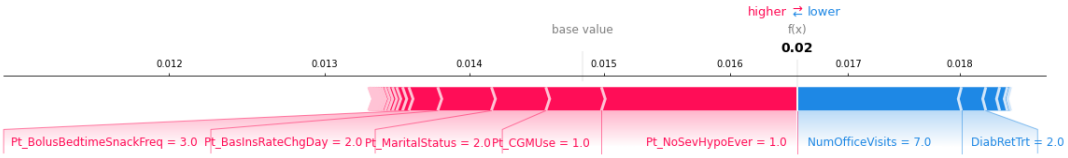

Figure 23. Local interpretation - SH prediction of a single male patient.

4.4.2. SH-Female Prediction Model Interpretation

Best results for SH event predictions in females were achieved with the LightGBM classifier. This section interprets the trained LightGBM classifier with female test data.

Global interpretation: Similar to the male prediction model, individuals with negative Shapley values show possible SH events, and positive Shapley values describe non-SH events in Figure 25. High feature values are in red, low values are in blue, and missing values are represented by gray. In general, individuals

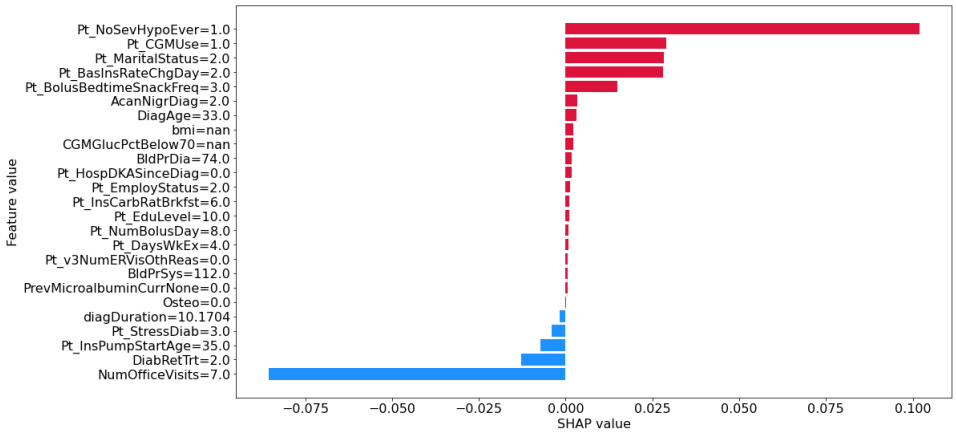

Figure 24. Detailed illustration of Figure 23.

who have ever experienced SH episodes are at high risk of having SH events again. Having foot ulcer, high duration of T1D, high number of visits to the healthcare provider, high breakfast insulin to carb ratio, a patient describes his general health as poor, lower height, higher weight, less number of long-acting insulin units and having DKA occurrences in the last 12 months shows a significant impact on the model to decide the patient with risk of having SH events. Furthermore, data on pregnancy status and menstrual cycle showed lower Shapley values. Females who are not pregnant and have irregular menstrual cycles influence the model's prediction.

Local interpretation: Figure 26 illustrates the interpretation of the model prediction of a sample in the female data set. The sample is a false negative case that the patient belongs to the SH event class since she had SH episodes within the previous 12 months, but the model predicted wrongly and classified her as a non-SH sample. A more detailed elaboration of Figure 26 can be seen in Figure 27.

The scale in Figure 26 represents the model predicted values (converted into probability) while the X-axis in Figure 27 represents the SHAP values.

The selected patient has never experienced an SH episode before ($Pt\_NoSevHypoEver$ = 1), has a low duration of diabetes ($diagDuration$ = 2.65), has a low breakfast insulin-to-carb ratio ($InsCarbRatBrkfst$ = 9) and described her general health as 'Good' ($Pt\_GenHealth$ = 2). As mentioned in the global interpretation section, these feature values have a negative impact on the model. Only the foot ulcer feature provides a visible impact to the model to classify this sample as a positive case. The Shapley value, 0.03, is greater than the base value, 0.022, and the model classifies this sample as a patient without SH events.

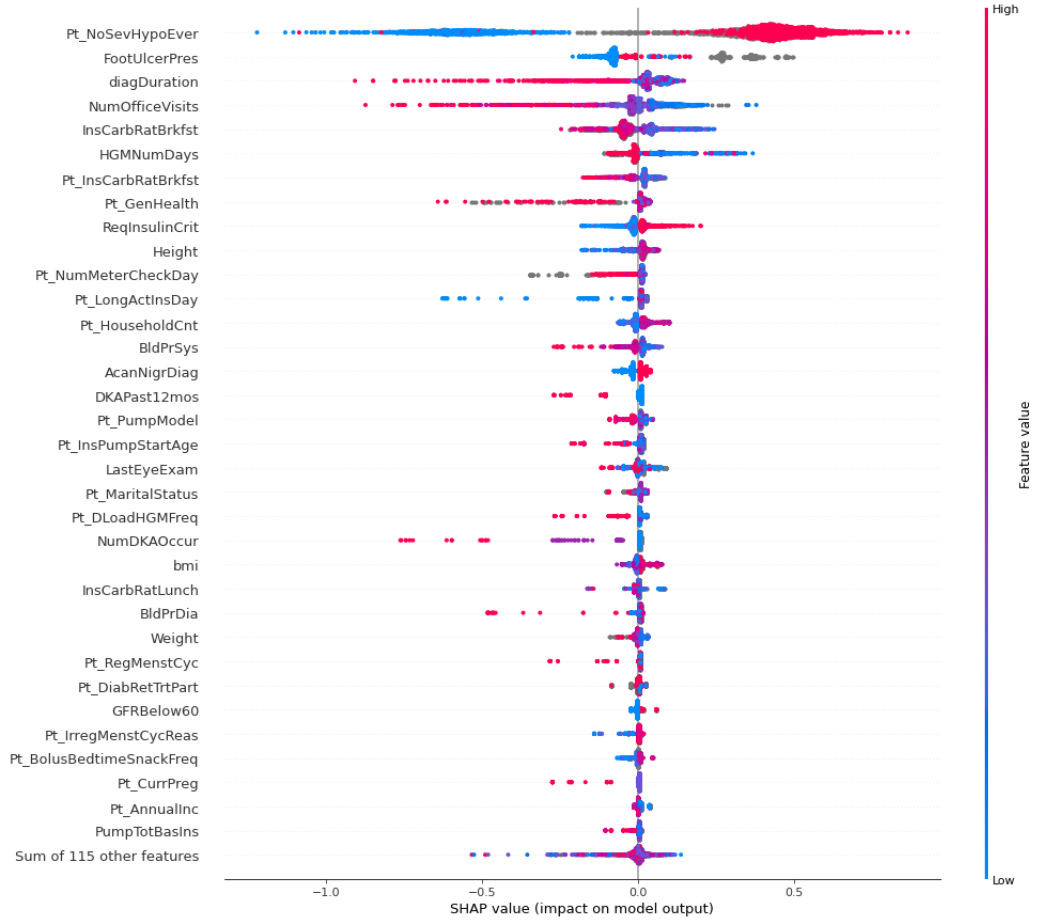

Figure 25. Global interpretation - SH prediction of female population.

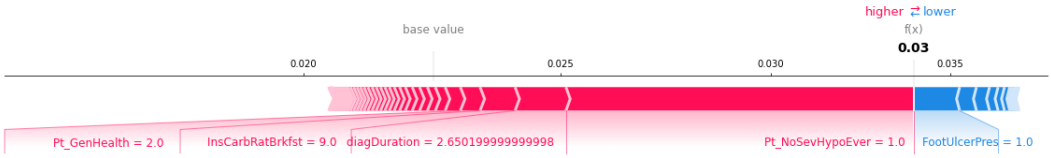

Figure 26. Local interpretation - SH prediction of a single female patient.

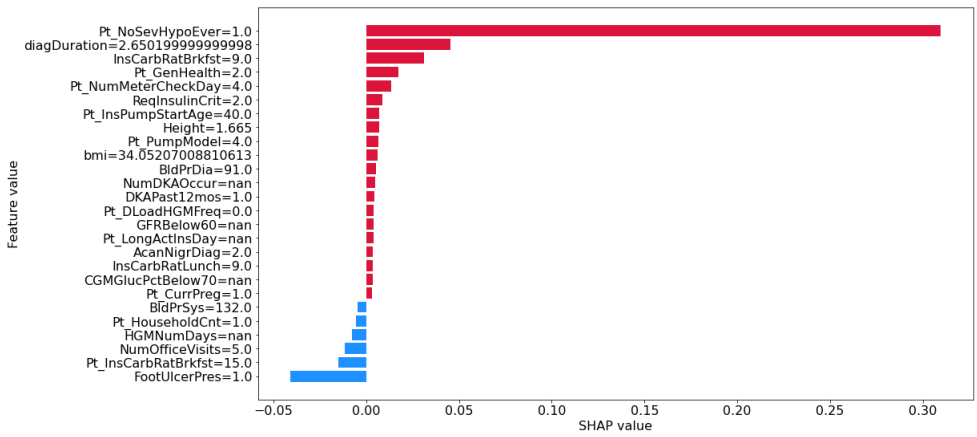

Figure 27. Detailed illustration of Figure 26.

4.4.3. DKA Prediction Model Interpretation

The best results were achieved with the LightGBM classification, to predict possible occurrences of DKA in T1D patients. This section contains the model explanation using the test data set.

Global interpretation: Negative SHAP values of the information dense summary plot in Figure 28 indicate possible DKA events, and positive SHAP values indicate non-DKA events. Patients' previous DKA history (ever had DKA in their life) strongly influences model prediction, where patients with DKA history have a high risk of having DKA again. Low weight, low duration of diabetes, low annual income, low mean HbA1c values, diagnosis with T1D at a young age, low BMI, not having private insurance, high mean blood glucose, high pulse rate, a considerable number of visits to a health provider, patients who mostly check ketones when blood sugar is high, patients defined their general health as 'poor', patients with diabetic peripheral neuropathy and with foot ulcer are more likely to encounter episodes of DKA within the next 12 months.

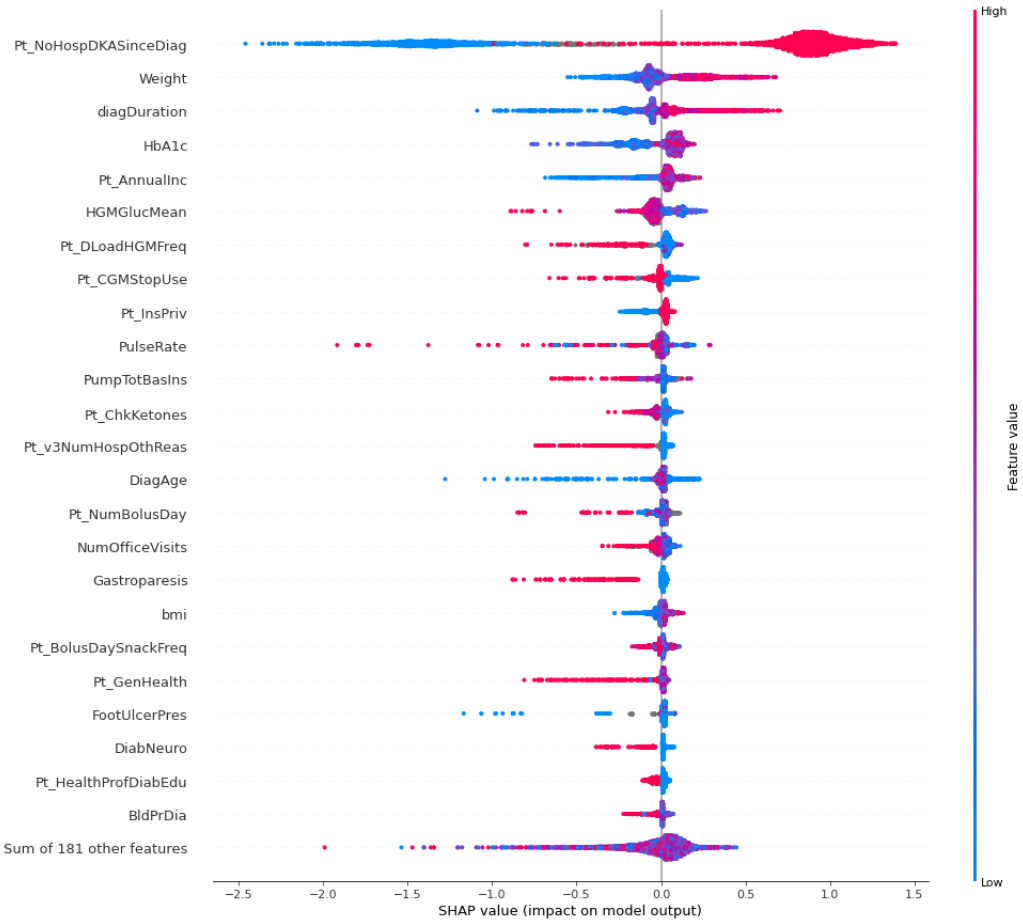

Figure 28. Global interpretation - DKA Prediction.

Local interpretation: Figure 29 visualizes the interpretation of the true positive sample in the test dataset, where the patient had DKA events in the last 12

months, and the model correctly classifies him/her as a patient with a DKA event. A more detailed illustration of Figure 29 can be seen in Figure 30.

Similar to the other model interpretation plots, the scale in Figure 29 represents the model predicted values (converted into probability) while the X-axis in Figure 30 represents the SHAP values.

The output SHAP values that are less than the base values will be classified as DKA events (positive decision) and values greater than the base values will be considered non-DKA event samples. This patient has a history of DKA events ($Pt\_NoHospDKASinceDiag = 0$), weighed 59.2 kg, has a mean blood glucose of 306 mg / dL and has been living with diabetes for 16 years. These reasons drive the model to decide that this patient is an individual with a risk of having DKA events within the next 12 months. Furthermore, the history of DKA events shows the highest impact on the decision. Although he has a high annual income in the range of $100,000-200,000$ and stops using continuous glucose monitoring ($Pt\_CGMStopUse = 0$), feature values in blue outperform these features and provide a positive impact on the model. The predicted SHAP value (0.0) is less than the base value (0.01). Therefore, the model classifies this sample as a patient with DKA.

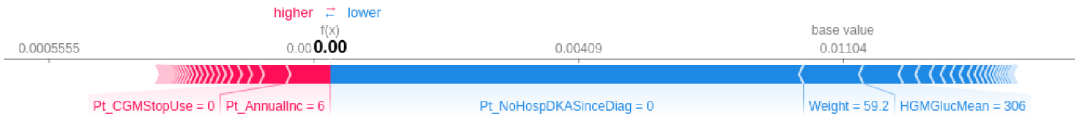

Figure 29. Local interpretation - DKA Prediction.

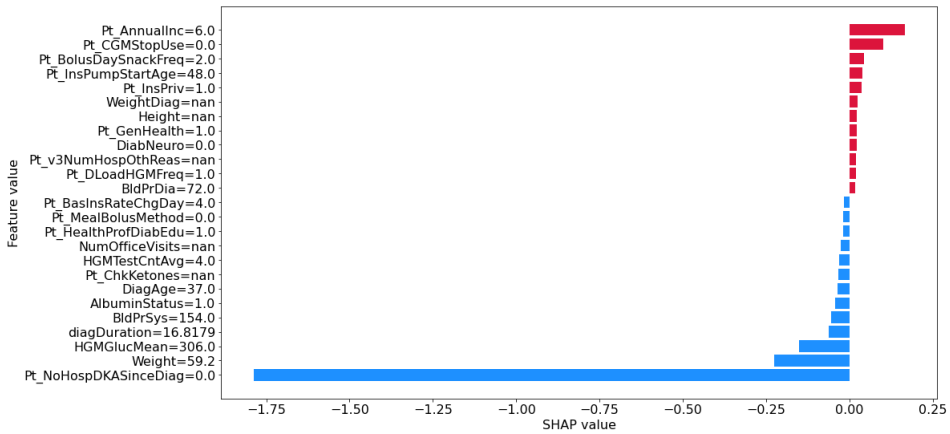

Figure 30. Detailed illustration of Figure 29.

4.5. Decision Support System

The implemented models were used to build a decision support system for health care personnel. The main objective of this system is to identify patients who are at high risk of developing SH or DKA events in the future. An algorithm was developed to calculate the high, moderate, and low priority decision boundaries using a training

dataset, model prediction probabilities, allocated cost for follow-ups/treatments, and estimated treatment cost per person. All the experiments in this section were conducted with the final DKA model.

We can define a method to prioritize patients based on model predicted probability. Figure 31 describes a normalized histogram of probabilities predicted by the DKA model. The actual values are represented by color, where positive DKA cases are represented in black, and negative DKA cases are colored yellow.

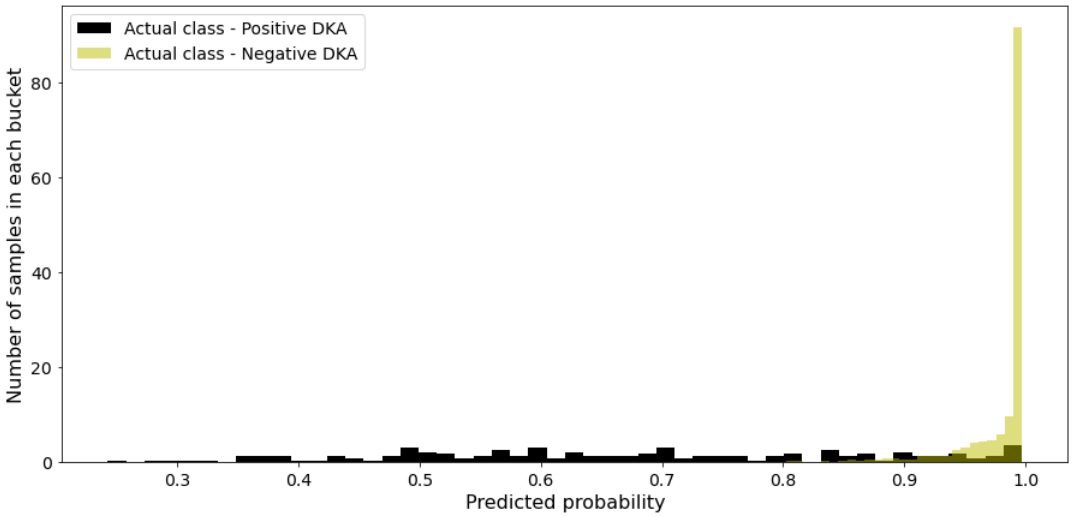

Figure 31. Histogram of predicted probabilities.

The base of this decision support system is the model prediction probabilities. We use two thresholds to categorize patients into risk groups where these thresholds can be defined using allocated costs for diabetes management. These thresholds can adjusted according to the situation. The risk threshold calculation algorithm is given in Algorithm 1, where it only considers the variability of high risk threshold (T_1). Based on the cost allocation, moderate risk threshold (T_2) can be derived manually.

Figure 32 used these thresholds and predicted probabilities to illustrate the system. Each point in Figure 32 represents a patient in the data set, and the Y-axis represents the probability predicted by the model. Point color indicates the actual label data, where yellow represents patients with non-DKA episodes and black represents patients with DKA. Defined thresholds are used to separate patients into risk categories where the area in red color indicates the high-risk patients, yellow indicates moderate-risk patients, and green indicates people with a low risk of developing DKA events in the next 12 months.

However, there are positive cases in both moderate and low-risk groups and negative cases in the high-risk group. Health care personnel can review the reasons for an individual's predicted probability using the local interpretation method described in Section 4.4.3.

We carried out an analysis with the positive DKA (black) individuals with negative DKA patients (yellow) in the high-risk group. Figure 33 illustrates the eight most important feature contributions of positive cases (black) in high-risk area and Figure 34 represents eight most important features contribute to negative cases (yellow) in high-risk area. Negative DKA (yellow) patients in the high-risk category show

Algorithm 1. Algorithm to calculate high risk threshold

Data: number of patients (N), predicted probabilities (P), allocated cost (C), expenses per patient (E)

Function `CalculateHighRiskThreshold(`P, C, E`)`**:**
  Init: High risk threshold (T_1), $margin$
  Treatable patient count $(n) \longleftarrow int(C/E)$
  if $n >= N$ **then**
    $T_1 \longleftarrow$ 1,0
  else
    Sort: P
    $T_1 \longleftarrow$ Value corresponds to n^{th} index in P + $margin$
  end
  return T_1
End Function

similar background to the positive DKA patients. Both figures provide similar feature variations, where low diabetes duration, low annual income, and low rate of general health were common and those features show a high impact on the prediction values. However, the presence of cardiomyopathy and the high number of hospitalization for other reasons shows a high impact on true positive patients while those features were not identified as important in false positive patients. In addition, patients with low BMI, high diastolic blood pressure ($BldPrDia$) values, and a high number of insulin boluses per day ($Pt\_NumBolusDay$) indicate a high impact on false positive predictions, and those two features were not identified as important features in true positive cases. The model identifies patients with these feature values or a combination of these features as patients with a high risk to develop DKA events within the next 12 months.

Furthermore, The false positive cases have low SHAP values compared to the true positive cases where the SHAP values of Figure 34 range from -1 to 1 while SHAP values of Figure 33 range from -2 to 2. In Figure 32, it is noticeable that the negative DKA cases are borderline predictions that model prediction probability values are varying between a 0.6 and the high threshold whereas positive cases are varying from 0.2 to a high threshold. So, the system considers these negative DKA patients as individuals that have a high risk of developing DKA events.

Similarly, decision support systems for SH prediction models were developed to identify the priority groups.

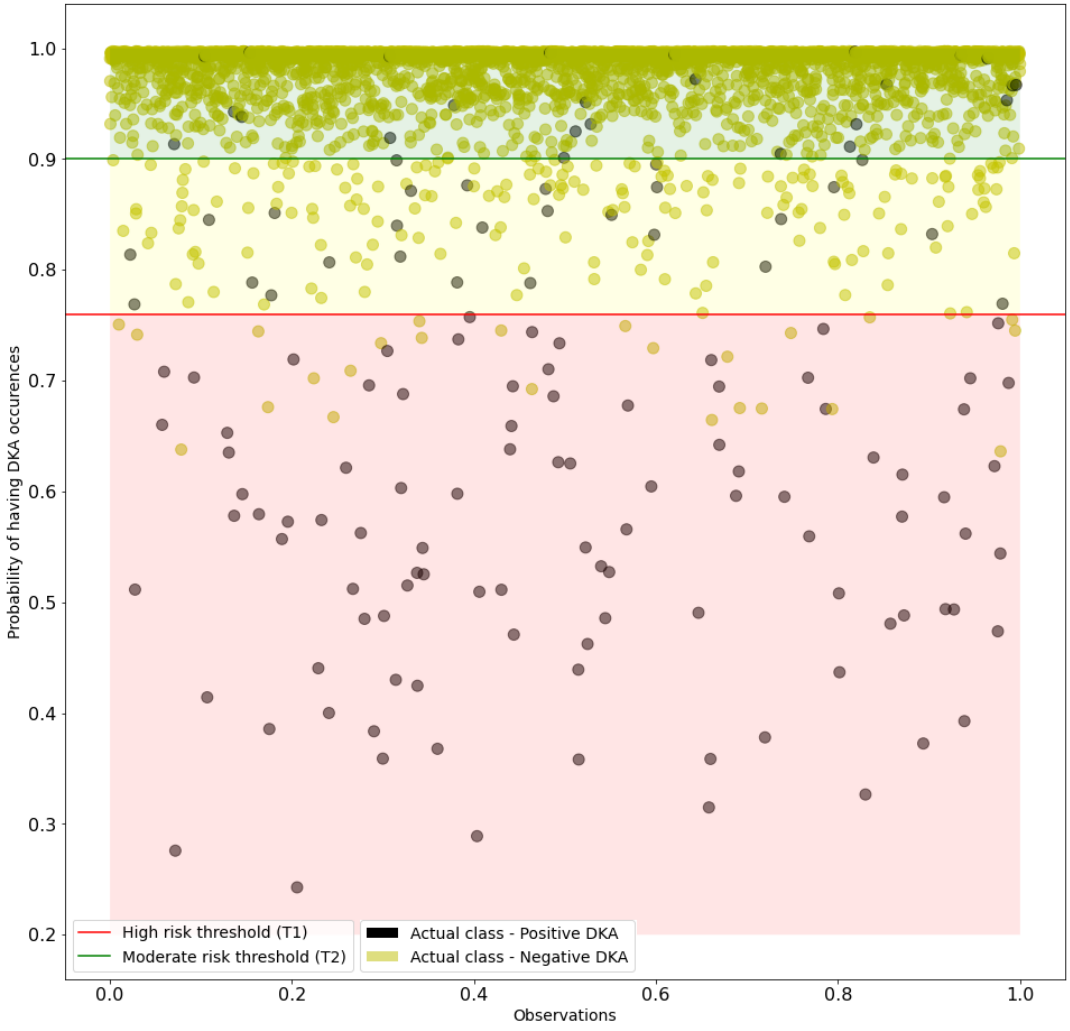

Figure 32. Decision support system.

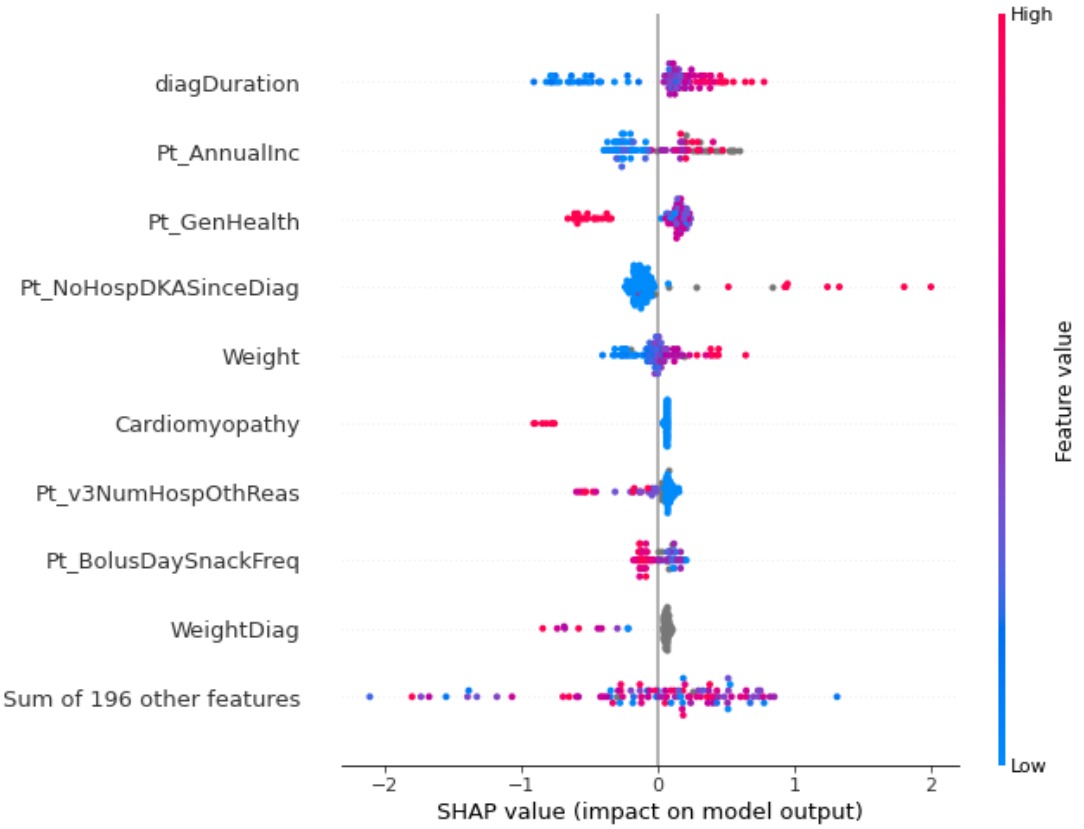

Figure 33. True positive cases in high risk area of Figure 32.

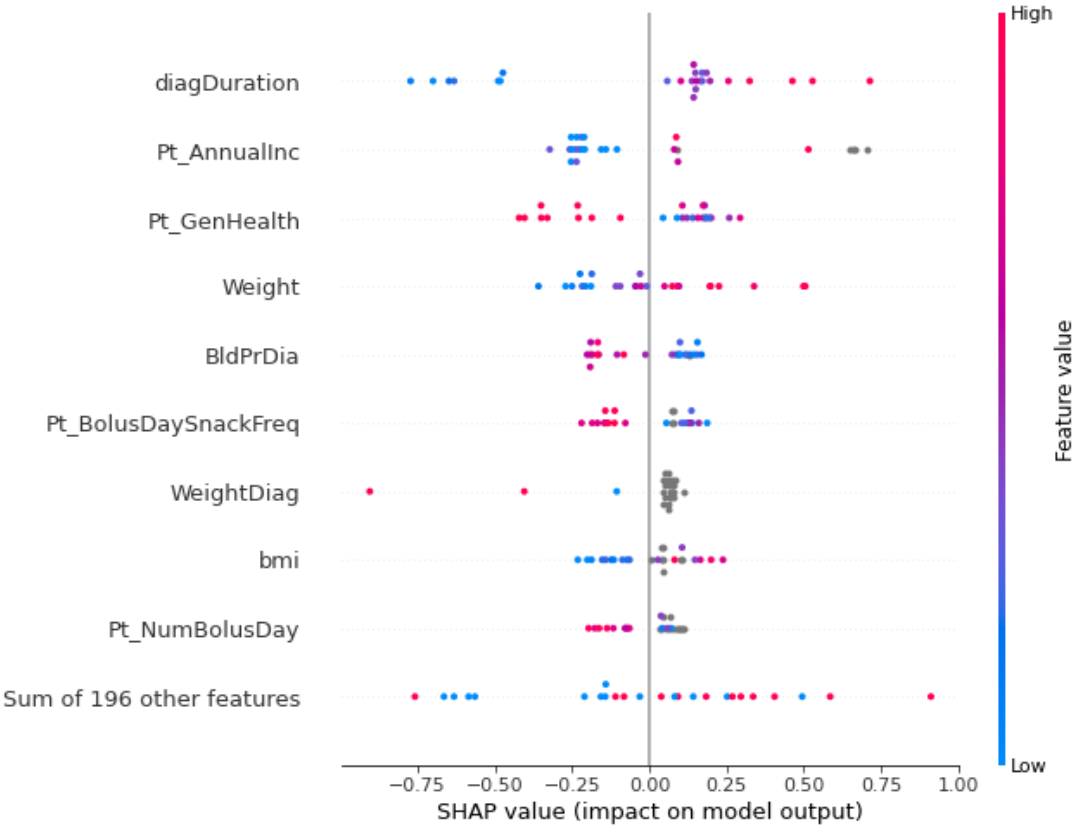

Figure 34. False positive cases in high risk area of Figure 32.

5. DISCUSSION

This study was carried out to develop a solution for the early identification of SH and DKA events in T1D patients using T1D Exchange registry data. Two different models were developed to predict SH events, one for males and another for the female study group. A single model was built for DKA prediction, including both male and females. Each predictive model is able to interpret the model itself and its individual results. Global interpretation also demonstrates risk factors of SH/DKA. Finally, a decision support system was introduced to identify priority patient groups.

Ten different models were trained for both SH and DKA prediction: LightGBM, XGBoost, AdaBoost, Random Forest, Decision Tree, Logistic regression, Linear discriminant analysis, Gaussian Naïve Bayes, SVM, and KNN classifier to select the best model. Both LightGBM and XGBoost classifiers allowed to have missing values in the training dataset. Hence, we trained both models with and without missing values in training dataset. LightGBM and XGBoost models, trained with missing values, outperform all the other models when predicting both SH and DKA events. However, the traditional machine learning models, KNN, SVM, Logistic regression, and Gaussian Naïve Bayes showed relatively poor performances.

It is common to have data imbalance in real-world medical data where the target class has uneven distributions of samples. In our case, many observations were available for the negative classes, and a comparatively less number of observations were in the positive classes. However, it is more crucial to identify the patients with SH/DKA events than non-SH/non-DKA patients. All three models have false positives where the model identified non-SH/non-DKA patients as positive cases. The models may identify the SH or DKA risk factors in these FP cases even though patients did not have SH or DKA occurrences in the last 12 months.

5.1. Factors Associated with the Predictive Models

The initial step of this study was to use the associations found in the replication study [54] to predict SH and DKA occurrences. The authors have used the logistic regression technique to find the association between 13 selected factors and event occurrences in the past 12 months. While they have used statistical methods in their work, we tried to use machine learning models and their findings to predict possible SH and DKA occurrences. We used these 13 features as baseline predictors to build predictive models, and both models showed poor performance in identifying SH and DKA events. Therefore, we concluded that these baseline features alone are not enough to provide adequate models. Further, we were able to improve these predictive models after carrying out several experiments.

Introducing more features and building separate models for male and female groups improved the SH model performance significantly. However, some factors mentioned in the replication study showed a significant influence on the improved models. The replication study found a strong association of SH events with a longer duration of diabetes. The SH-Female predictive model showed a positive prediction influence with high diabetes duration while the SH-Male predictive model showed a negative impact. Low household income showed an association with SH events in [54]. Similarly,

the SH-Female model showed a high risk of developing SH events in patients with low household incomes. Furthermore, BMI feature showed an influence in the SH-Female prediction model, where [54] states SH events are independent of BMI. Other associated factors found by [54] did not reveal a significant impact on both SH model predictions.

A considerable number of factors mentioned in the replication study showed an influence on DKA model prediction. Replication study showed lower age, HbA1c levels, females, BMI, low education, low income, no private insurance, non-Hispanic black/Hispanic ethnicity, current smoking status are associated with higher DKA frequencies. The built DKA predictive model identified low annual income, lower levels of HbA1c, low BMI, lack of private insurance coverage, lower age, and ethnicity as characteristics that strongly influenced the prediction of DKA events. A replication study revealed the frequency of DKA was not significantly associated with the diabetic duration and insulin method. However, the final prediction model showed a strong influence on the prediction of positive DKA events in individuals with a shorter duration of diabetes.

Of the 13 factors mentioned in the replication paper, more than six were noticeable in the DKA prediction model, while only three features influence the SH models. In addition to these factors, having a history with SH and DKA events showed a high influence on the prediction of positive outcomes. All the models used the number of visits to a health provider and diastolic blood pressure data when predicting outcomes. Some common associations were found in both SH-Male and SH-Female models like marital status and basal insulin rate change. Moreover, diabetes-related stress, diabetic retinopathy, and acanthosis nigricans attributes were notable in the SH-Male prediction model while foot ulcer, general health, height, weight, DKA occurrences in past 12 months, pregnancy status, menstrual cycle data, and glucose monitoring data were significant in SH-Female model. Furthermore, the DKA prediction model identified weight, number of times patients check for ketones when blood sugar is high, pulse rate, foot ulcer, diabetic peripheral neuropathy, glucose monitoring data, basal insulin data, and general health as outstanding features for the prediction.

5.2. Decision Support System

Clinical decision support systems help to enhance the medical decisions taken by healthcare providers. In this work, we provide a clinical support system that can be used to identify individuals in the high-, moderate-, and low-risk groups of developing SH and DKA events. This system will facilitate precision medicine where it helps healthcare providers to identify high-risk patients, so they can prioritize patients and provide additional care.

With this system, healthcare providers do not have to focus on every T1D patient; they can focus on high-risk patients first. Both of these outcomes are associated with high medical expenses. The decision support system was built based on SH or DKA-related cost allocation in medical follow-ups and treatments where it identifies the trade-off between the allocated budget for SH/DKA events and the number of patients in the high-risk group. Higher the amount allocated to these outcomes, higher the count of treatable patients. The introduced decision support system helps to provide

an optimal allocation of medical resources and medical expenses by focusing on the group that needs care.

This system will help to leverage health care quality because the system provides information on why the patient got categorized into a high-risk group. This helps healthcare personnel to have a better understanding of the individual patient they are working with. Furthermore, this will help to provide patient-specific care where the individuals in identified high-risk group may require additional visits to a healthcare provider, changes in their treatments, and more attention.

Risks can be productively assessed because the system allows one to make informed and stable decisions. Healthcare personnel are aware of the reasons for categorizing an individual into a high-risk group. This results in providing T1D patients with proper diabetes management and leads them to have a better quality of life.

5.3. Limitations and Future Work

This study was carried out with the T1D exchange registry enrollment dataset with patients aged 26 years or older and had a T1D duration of at least two years to predict possible SH and DKA occurrences in the next 12 months. After this filtration, only 7156 patient data were available to develop the models. One of the main limitations of this study is that we focus only on adult patients with T1D. However, the literature showed that DKA events are much more common among young children and adolescents, frequency of DKA events tends to decrease with the age [49, 50]. Further analysis and improvements to the DKA prediction model can be done using all T1D patients in the dataset. Furthermore, the built models were trained using the T1D exchange registry dataset that contains data of T1D patients in the United States. There can be differences in patients in different countries. Hence to get a better understanding and test the generalizability of the models, further model testing needs to be carried out with data from different countries.

Both models were trained with an imbalanced dataset where the majority of patients did not have any SH or DKA occurrences in the past 12 months. Some missing values in the data set were handled by assuming that the patients filled out the questionnaire correctly.

The SHAP library was used to interpret model outcomes and assumes feature independence for the computation of Shapley values. In most real-world datasets, features are not statistically independent. Even though SHAP plots discovered similar characteristics as in literature, due to this limitation, the calculated Shapley values can suffer from uncertainty. Hence, checking the relevance of a model's global and local interpretations is an essential task. Knowledge of domain expertise can be used to verify the validity of the interpretations of the models concluded in this study.

This study focused on developing predictive models that can forecast SH and DKA events. Future work could focus on using predictive models and decision support systems to implement a solid system that healthcare professionals can use in their daily practice. Finally, introduced decision support system requires a more knowledgeable way to calculate the threshold value that separates the moderate-risk and low-risk groups.

6. CONCLUSION

Type 1 diabetes is a chronic disease that requires close attention and proper management to keep it under control. SH and DKA are major life-threatening complications for people with type 1 diabetes. This is an important research area in the field of type 1 diabetes care, where numerous studies have been carried out to find associated risk factors and for early prediction of these outcomes. This thesis is motivated by Weinstock et al. [54]'s research study that discover the risk factors associated with SH and DKA in T1D adults. We used the findings of the original paper [54] as baseline factors to develop predictive models that forecast possible SH and DKA occurrences in the next 12 months. However, the baseline factors alone did not provide enough information to train a performant model. Therefore, more features were used to further improve these predictive models. Final SH-Female model predicted possible SH events in female population with 78% balanced accuracy while SH-Male prediction model achieved higher balanced accuracy of 86% with 100% accuracy for identifying positive cases. DKA model achieved 82% balanced accuracy with 77% F1 score. The final models highlighted some of the baseline factors in the original study when predicting the outcomes.

AI in healthcare values both accuracy and interpretability. Many high-performing ML algorithms are still considered black boxes, where it is hard to understand the model's inner mechanism after they have been trained. Under the General Data Protection Regulation (GDPR), automated processes that use personal data must be able to explain the reasons behind the outcome. To compliant with GDPR it is required to explain how the final models take the decision [88]. However, an explanation for model outcomes is essential to trust the model, especially in sensitive areas like the medical field. We used the SHAP library to explain the trained model itself and its individual outcomes. This helps to identify biases, avoid mistakes, and gain a reasonable understanding of the final models.

Model explanations are even more important when models are used to build a decision support system where it helps understanding the factors behind the decision. The decision support system implemented with this study can identify high-risk, moderate-risk, and low-risk patient groups. This facilitates precision medicine and enhances the medical decisions on target groups since healthcare personnel can identify and focus on high-risk patients in advance. Furthermore, this will reduce the number of patients to focus on in the clinical center and help to reduce the cost associated with SH and DKA through efficient resource management.

The developed system aims to improve the quality of life of T1D patients by identifying SH and DKA events within a 12-month forecast window. This will help both healthcare personnel and patient to manage diabetes in an efficient way.

7. REFERENCES

[1] Klonoff D.C. (2015) Precision medicine for managing diabetes. Journal of Diabetes Science and Technology 9, pp. 3–7. URL: https://doi.org/10.1177/1932296814563643, pMID: 25550409.

[2] Clish C.B. (2015) Metabolomics: an emerging but powerful tool for precision medicine. Molecular Case Studies 1, p. a000588.

[3] Facts & figures. URL:https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html. Accessed 08.02.2022.

[4] Graziella B., Picariello R., Petrelli A., Panero F., Costa G., Cavallo-Perin P., Demaria M. & Gnavi R. (2012) Direct costs in diabetic and non diabetic people: the population-based turin study, italy. Nutrition, Metabolism and Cardiovascular Diseases 22, pp. 684–690.

[5] Association A.D. (2018) Economic costs of diabetes in the us in 2017. Diabetes care 41, pp. 917–928.

[6] Gosmanov A.R., Gosmanova E.O. & Kitabchi A.E. (2021) Hyperglycemic crises:diabetic ketoacidosis and hyperglycemic hyperosmolar state. Endotext [Internet] .

[7] Lacy M.E., Gilsanz P., Eng C., Beeri M.S., Karter A.J. & Whitmer R.A. (2020) Severe hypoglycemia and cognitive function in older adults with type 1 diabetes: the study of longevity in diabetes (solid). Diabetes Care 43, pp. 541–548.

[8] Darpit D., DeSalvo D.J., Balakrishna H., McKay S., Shenoy A., Chester K.J., Lawley M. & Erraguntla M. (2021) Feature-based machine learning model for real-time hypoglycemia prediction. Journal of Diabetes Science and Technology 15, pp. 842–855.

[9] Aliberti A., Pupillo I., Stefano T., Macii E., Di Cataldo S., Patti E. & Acquaviva A. (2019) A multi-patient data-driven approach to blood glucose prediction. IEEE Access 7, pp. 69311–69325.

[10] Vakkuri V., Kemell K.K. & Abrahamsson P. (2020) Eccola-a method for implementing ethically aligned ai systems. In: 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, pp. 195–204.

[11] Beck R.W., Tamborlane W.V., Bergenstal R.M., Miller K.M., DuBose S.N., Hall C.A. & for the T1D Exchange Clinic Network (2012) The T1D Exchange Clinic Registry. The Journal of Clinical Endocrinology & Metabolism 97, pp. 4383–4389. URL: https://doi.org/10.1210/jc.2012-1561.

[12] HTx Project | Next Generation Health Technology Assessment. URL:https://www.htx-h2020.eu/. Accessed 12.02.2022.

[13] Ahmed A.M. (2002) History of diabetes mellitus. Saudi medical journal 23, pp. 373–378.

[14] Most diabetic country 2019 | statista_2022. URL: https://www.statista.com/statistics/281082/countries-with-highest-number-of-diabetics/. Accessed 10.02.2022.

[15] What is diabetes 2022. URL: https://www.cdc.gov/diabetes/basics/diabetes.html. Accessed 15.02.2022.

[16] Diabetes now affects one in 10 adults worldwide. URL: https://www.idf.org/news/240:diabetes-now-affects-one-in-10-adults-worldwide.html. Accessed 15.02.2022.

[17] Asif M. (2014) The prevention and control the type-2 diabetes by changing lifestyle and dietary pattern. Journal of education and health promotion 3.

[18] What is type 1 diabetes. URL: https://www.cdc.gov/diabetes/basics/what-is-type-1-diabetes.html. Accessed 16.02.2022.

[19] Mobasseri M., Shirmohammadi M., Amiri T., Vahed N., Fard H.H. & Ghojazadeh M. (2020) Prevalence and incidence of type 1 diabetes in the world: a systematic review and meta-analysis. Health promotion perspectives 10, p. 98.

[20] Atkinson M.A., Eisenbarth G.S. & Michels A.W. (2014) Type 1 diabetes. The Lancet 383, pp. 69–82.

[21] Distiller L.A. (2014) Why do some patients with type 1 diabetes live so long? World journal of Diabetes 5, p. 282.

[22] How Diabetes Affects the Whole Family. URL: https://www.byramhealthcare.com/blogs/how-diabetes-affects-the-whole-family. Accessed 21.02.2022.

[23] Dall T.M., Mann S.E., Zhang Y., Quick W.W., Seifert R.F., Martin J., Huang E.A. & Shiping Z. (2009) Distinguishing the economic costs associated with type 1 and type 2 diabetes. Population health management 12, pp. 103–110.

[24] Tao B., Pietropaolo M., Atkinson M., Schatz D. & Taylor D. (2010) Estimating the cost of type 1 diabetes in the us: a propensity score matching method. PloS one 5, p. e11501.

[25] Seaquist E.R., Anderson J., Childs B., Cryer P., Dagogo-Jack S., Fish L., Heller S.R., Rodriguez H., Rosenzweig J. & Vigersky R. (2013) Hypoglycemia and diabetes: a report of a workgroup of the american diabetes association and the endocrine society. The Journal of Clinical Endocrinology & Metabolism 98, pp. 1845–1859.

[26] Cryer P.E., Fisher J.N. & Shamoon H. (1994) Hypoglycemia. Diabetes care 17, pp. 734–755.

[27] Jensen M.H., Dethlefsen C., Hejlesen O. & Vestergaard P. (2020) Association of severe hypoglycemia with mortality for people with diabetes mellitus during a 20-year follow-up in denmark: a cohort study. Acta Diabetologica 57, pp. 549–558.

[28] Association A.D. (2014) 6. Glycemic Targets. Diabetes Care 38, pp. S33–S40. URL: https://doi.org/10.2337/dc15-S009.

[29] Chiang J.L., Kirkman M.S., Laffel L.M., Peters A.L. & Authors T..D.S. (2014) Type 1 diabetes through the life span: a position statement of the american diabetes association. Diabetes care 37, pp. 2034–2054.

[30] Weinstock R.S., DuBose S.N., Bergenstal R.M., Chaytor N.S., Peterson C., Olson B.A., Munshi M.N., Perrin A.J., Miller K.M., Beck R.W., Liljenquist D.R., Aleppo G., Buse J.B., Kruger D., Bhargava A., Goland R.S., Edelen R.C., Pratley R.E., Peters A.L., Rodriguez H., Ahmann A.J., Lock J.P., Garg S.K., Rickels M.R., Hirsch I.B. & for the T1D Exchange Severe Hypoglycemia in Older Adults With Type 1 Diabetes Study Group (2015) Risk Factors Associated With Severe Hypoglycemia in Older Adults With Type 1 Diabetes. Diabetes Care 39, pp. 603–610. URL: https://doi.org/10.2337/dc15-1426.

[31] Allen C., LeCaire T., Palta M., Daniels K., Meredith M., D'Alessio D.J. & Project W.D.R. (2001) Risk factors for frequent and severe hypoglycemia in type 1 diabetes. Diabetes care 24, pp. 1878–1881.

[32] Giorda C.B., Ozzello A., Gentile S., Aglialoro A., Chiambretti A., Baccetti F., Gentile F.M., Lucisano G., Nicolucci A. & Rossi M.C. (2015) Incidence and risk factors for severe and symptomatic hypoglycemia in type 1 diabetes. results of the hypos-1 study. Acta diabetologica 52, pp. 845–853.

[33] Pedersen-Bjergaard U., Pramming S., Heller S.R., Wallace T.M., Rasmussen Å.K., Jørgensen H.V., Matthews D.R., Hougaard P. & Thorsteinsson B. (2004) Severe hypoglycaemia in 1076 adult patients with type 1 diabetes: influence of risk markers and selection. Diabetes/metabolism research and reviews 20, pp. 479–486.

[34] Sämann A., Lehmann T., Heller T., Müller N., Hartmann P., Wolf G.B. & Müller U.A. (2013) A retrospective study on the incidence and risk factors of severe hypoglycemia in primary care. Family practice 30, pp. 290–293.

[35] Ringholm L., Pedersen-Bjergaard U., Thorsteinsson B., Damm P. & Mathiesen E. (2012) Hypoglycaemia during pregnancy in women with type 1 diabetes. Diabetic Medicine 29, pp. 558–566.

[36] Zhang Y. (2008) Predicting occurrences of acute hypoglycemia during insulin therapy in the intensive care unit. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, pp. 3297–3300.

[37] Georga E.I., Protopappas V.C., Ardigò D., Polyzos D. & Fotiadis D.I. (2013) A glucose model based on support vector regression for the prediction of

hypoglycemic events under free-living conditions. Diabetes technology & therapeutics 15, pp. 634–643.

[38] Ruan Y., Bellot A., Moysova Z., Tan G.D., Lumb A., Davies J., Van Der Schaar M. & Rea R. (2020) Predicting the risk of inpatient hypoglycemia with machine learning using electronic health records. Diabetes care 43, pp. 1504–1511.

[39] Cox D.J., Gonder-Frederick L., Ritterband L., Clarke W. & Kovatchev B.P. (2007) Prediction of severe hypoglycemia. Diabetes Care 30, pp. 1370–1373.

[40] Schroeder E.B., Xu S., Goodrich G.K., Nichols G.A., O'Connor P.J. & Steiner J.F. (2017) Predicting the 6-month risk of severe hypoglycemia among adults with diabetes: development and external validation of a prediction model. Journal of Diabetes and its Complications 31, pp. 1158–1163.

[41] Reddy R., Resalat N., Wilson L.M., Castle J.R., El Youssef J. & Jacobs P.G. (2019) Prediction of hypoglycemia during aerobic exercise in adults with type 1 diabetes. Journal of diabetes science and technology 13, pp. 919–927.

[42] Johnson S., Cooper M., Davis E. & Jones T. (2013) Hypoglycaemia, fear of hypoglycaemia and quality of life in children with type 1 diabetes and their parents. Diabetic medicine 30, pp. 1126–1131.

[43] Mujahid O., Contreras I. & Vehi J. (2021) Machine learning techniques for hypoglycemia prediction: Trends and challenges. Sensors 21, p. 546.

[44] VanItallie T.B. & Nufert T.H. (2003) Ketones: metabolism's ugly duckling. Nutrition reviews 61, pp. 327–341.

[45] Abbas Q., Arbab S., Ul Haque A. & Humayun K.N. (2018) Spectrum of complications of severe dka in children in pediatric intensive care unit. Pakistan journal of medical sciences 34, p. 106.

[46] Umpierrez G.E. & Kitabchi A.E. (2003) Diabetic ketoacidosis. Treatments in endocrinology 2, pp. 95–108.

[47] Dhatariya K., Skedgel C. & Fordham R. (2017) The cost of treating diabetic ketoacidosis in the uk: a national survey of hospital resource use. Diabetic Medicine 34, pp. 1361–1366.

[48] Misra S. & Oliver N.S. (2015) Diabetic ketoacidosis in adults. BMJ 351.

[49] Farsani S.F., Brodovicz K., Soleymanlou N., Marquard J., Wissinger E. & Maiese B.A. (2017) Incidence and prevalence of diabetic ketoacidosis (dka) among adults with type 1 diabetes mellitus (t1d): a systematic literature review. BMJ open 7, p. e016587.

[50] Ehrmann D., Kulzer B., Roos T., Haak T., Al-Khatib M. & Hermanns N. (2020) Risk factors and prevention strategies for diabetic ketoacidosis in people with established type 1 diabetes. The Lancet Diabetes & Endocrinology 8, pp. 436–446.

[51] Al-Obaidi A.H., Alidrisi H.A. & Mansour A.A. (2019) Precipitating factors for diabetic ketoacidosis among patients with type 1 diabetes mellitus: the effect of socioeconomic status. Dubai Diabetes and Endocrinology Journal 25, pp. 52–60.

[52] Gilhotra Y. & Porter P. (2007) Predicting diabetic ketoacidosis in children by measuring end-tidal co2 via non-invasive nasal capnography. Journal of Paediatrics and Child Health 43, pp. 677–680.

[53] Li L., Lee C.C., Zhou F.L., Molony C., Doder Z., Zalmover E., Sharma K., Juhaeri J. & Wu C. (2021) Performance assessment of different machine learning approaches in predicting diabetic ketoacidosis in adults with type 1 diabetes using electronic health records data. Pharmacoepidemiology and drug safety 30, pp. 610–618.

[54] Weinstock R.S., Xing D., Maahs D.M., Michels A., Rickels M.R., Peters A.L., Bergenstal R.M., Harris B., DuBose S.N., Miller K.M. et al. (2013) Severe hypoglycemia and diabetic ketoacidosis in adults with type 1 diabetes: results from the t1d exchange clinic registry. The Journal of Clinical Endocrinology & Metabolism 98, pp. 3411–3419.

[55] National Diabetes Statistics Report 2020. URL: `https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf`.

[56] Brownlee J. (2016) Machine learning algorithms from scratch with Python. Machine Learning Mastery.

[57] Van Engelen J.E. & Hoos H.H. (2020) A survey on semi-supervised learning. Machine Learning 109, pp. 373–440.

[58] Sutton R.S. & Barto A.G. (2018) Reinforcement learning: An introduction. MIT press.

[59] Kotsiantis S.B., Kanellopoulos D. & Pintelas P.E. (2006) Data preprocessing for supervised leaning. International journal of computer science 1, pp. 111–117.

[60] Aggarwal C.C. & Yu P.S. (2001) Outlier detection for high dimensional data. In: Proceedings of the 2001 ACM SIGMOD international conference on Management of data, pp. 37–46.

[61] Alimohammadi H. & Chen S.N. (2022) Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis. Expert Systems with Applications 191, p. 116371.

[62] Kannan K.S., Manoj K. & Arumugam S. (2015) Labeling methods for identifying outliers. International Journal of Statistics and Systems 10, pp. 231–238.

[63] Bennett D.A. (2001) How can i deal with missing data in my study? Australian and New Zealand journal of public health 25, pp. 464–469.

[64] How to handle missing data2022. URL: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4. Accessed 05.03.2022.

[65] Acuna E. & Rodriguez C. (2004) The treatment of missing values and its effect on classifier accuracy. In: Classification, clustering, and data mining applications, Springer, pp. 639–647.

[66] Nayak S., Misra B.B. & Behera H.S. (2014) Impact of data normalization on stock index forecasting. International Journal of Computer Information Systems and Industrial Management Applications 6, pp. 257–269.

[67] Abdi H. & Williams L.J. (2010) Principal component analysis. Wiley interdisciplinary reviews: computational statistics 2, pp. 433–459.

[68] LDA vs. PCA. URL: https://towardsai.net/p/data-science/lda-vs-pca. Accessed 05.03.2022.

[69] Machine learning model training: what it is and why it's important2022. URL: https://blog.dominodatalab.com/what-is-machine-learning-model-training. Accessed 13.03.2022.

[70] Ray S. (2019) A quick review of machine learning algorithms. In: 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), IEEE, pp. 35–39.

[71] Gupta B., Rawat A., Jain A., Arora A. & Dhami N. (2017) Analysis of various decision tree algorithms for classification in data mining. International Journal of Computer Applications 163, pp. 15–19.

[72] Qi Y. (2012) Random forest for bioinformatics. In: Ensemble machine learning, Springer, pp. 307–323.

[73] Joyce J. (2003) Bayes' theorem .

[74] Tharwat A., Gaber T., Ibrahim A. & Hassanien A.E. (2017) Linear discriminant analysis: A detailed tutorial. AI communications 30, pp. 169–190.

[75] Vezhnevets A. & Vezhnevets V. (2005) Modest adaboost-teaching adaboost to generalize better. In: Graphicon, vol. 12, vol. 12, pp. 987–997.

[76] Chen T. & Guestrin C. (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

[77] Xgboost documentation — xgboost 1.5.2 documentation2022. URL: https://xgboost.readthedocs.io/en/stable/. Accessed 17.03.2022.

[78] Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q. & Liu T.Y. (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30.

[79] Galar M., Fernandez A., Barrenechea E., Bustince H. & Herrera F. (2011) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, pp. 463–484.

[80] Brownlee J., A Gentle Introduction to Threshold-Moving for Imbalanced Classification. URL: https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/. Accessed 18.03.2022.

[81] Abdullah T.A., Zahid M.S.M. & Ali W. (2021) A review of interpretable ml in healthcare: Taxonomy, applications, challenges, and future directions. Symmetry 13, p. 2439.

[82] Ribeiro M.T., Singh S. & Guestrin C. (2016) "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

[83] Shrikumar A., Greenside P., Shcherbina A. & Kundaje A. (2016) Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 .

[84] Lundberg S.M. & Lee S.I. (2017) A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

[85] Explain your model with the shap values2022. URL: https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d. Accessed 17.03.2022.

[86] Fisher R.A. (1936) The use of multiple measurements in taxonomic problems. Annals of eugenics 7, pp. 179–188.

[87] Scikit-learn Machine learning in Python. URL: https://scikit-learn.org/stable/. Accessed 13.05.2022.

[88] General Data Protection Regulation. URL: https://gdpr.eu/Recital-71-Profiling. Accessed 19.05.2022.