Screening for type 2 diabetes from single lead electrocardiogram data:

A machine learning approach

Mariam Jabara, Division of Experimental Medicine, Faculty of Medicine and Health Sciences

McGill University, Montreal

December 2023

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of

Master of Science

First published on December 15th, 2024

Table of Contents

Abstract

BACKGROUND: Diabetes is one of the largest global health emergencies of the 21st century, with type 2 diabetes mellitus (T2DM) accounting for 90% of all diabetes worldwide. T2DM is typically diagnosed after a clinical event (e.g., a heart attack) or during routine screenings for patients of older age. However, T2DM can emerge much earlier in life and is often unrecognized, with its onset occurring at least 4-7 years before a clinical diagnosis. Consequently, there is a largely unmet clinical need for low-cost, non-invasive methods to screen for T2DM and pre-diabetes. Given that T2DM is a data-rich condition with various outcomes, artificial intelligence (AI) and machine learning (ML) models can be leveraged to learn from previous patient data and aid in early-detection and treatment of the disease. Electrocardiograms (ECGs) are data rich and when combined with AI may serve as an inexpensive, scalable, and effective method to screen for T2DM in the general population.

METHODS: We developed and validated a ML-based model to detect the presence of T2DM from Lead-I of a standard 12-Lead ECG using Extreme Gradient Boosting (XGBoost). The models were trained and tested on data from the CARTaGENE research project from CHU-St. Justine (n=7463). We conducted three experiments using time-series based, frequency-based and interval-based features of the ECG as input to the model.

RESULTS: The best models used frequency-based and interval-based features, achieving an AUROC of **0.784** and sensitivity of **0.882.** These findings suggest that it is possible to achieve similar results to state-of-the-art models without the deep learning (DL) approaches that they typically employ. Additionally, the models have demonstrated results on par with and exceeding gold standard screening strategies for T2DM: the CANRISK questionnaire and HbA1c testing.

CONCLUSION: We demonstrate the use of a ML-based model for the screening of T2DM using the highly efficient XGBoost algorithm. The proposed model accurately detects the presence or absence of the disease based on single-lead electrocardiogram data in patients from the CARTaGENE dataset. Overall, this may serve as an effective way to conduct population-level screening for T2DM, facilitating earlier detection and intervention of the disease and improving patient outcomes.

Resumé

CONTEXTE: Le diabète est l'une des plus grandes urgences sanitaires mondiales du 21ème siècle, le diabète sucré de type 2 DT2 représentant 90 % de tous les diabètes dans le monde (1). Le DT2 est généralement diagnostiqué après un événement clinique (par exemple, une crise cardiaque) ou lors de dépistages de routine pour les patients plus âgés. Cependant, le DT2 peut apparaître beaucoup plus tôt dans la vie et demeure souvent méconnu, son apparition survenant au moins 4 à 7 ans avant un diagnostic clinique (2). Par conséquent, il existe un besoin clinique largement insatisfait de méthodes peu coûteuses et non invasives pour dépister le DT2 et le prédiabète. Étant donné que le DT2 est une condition riche en données avec divers résultats, les modèles d'intelligence artificielle (IA) et d'apprentissage automatique (ML) peuvent être exploités pour apprendre des données antérieures des patients et aider à la détection précoce et au traitement de la maladie. Les électrocardiogrammes (ECG) sont riches en données et, lorsqu'ils sont combinés à l'IA, ils peuvent constituer une méthode peu coûteuse, évolutive et efficace pour dépister le DT2 dans la population générale.

MÉTHODES: Nous avons développé et validé un modèle basé sur l'apprentissage automatique (ML) pour détecter la présence de DT2 à partir de Lead-I d'un électrocardiogramme (ECG) standard à 12 dérivations à l'aide du modèle Extreme Gradient Boosting (XGBoost). Les modèles ont été entraînés et testés sur les données du projet de recherche CARTaGENE (CHU-Sainte-Justine) (n=7463). Nous avons mené trois expériences en utilisant des caractéristiques de l'ECG basées sur des séries chronologiques, des fréquences et des intervalles comme données d'entrée du modèle.

RÉSULTATS: Les meilleurs modèles utilisaient des fonctionnalités basées sur la fréquence et sur l'intervalle, atteignant Une aire sous la courbe ROC de 0,784 et une sensibilité de 0,882. Nos résultats suggèrent qu'il est possible d'obtenir des résultats similaires aux modèles de pointe sans les approches d'apprentissage en profondeur (DL) qu'ils utilisent généralement. De plus, nos modèles ont démontré des résultats comparables et supérieurs aux stratégies de dépistage de référence pour le DT2 : le questionnaire CANRISK et le test HbA1c.

CONCLUSION: Nous démontrons l'utilisation d'un modèle basé sur ML pour le dépistage du DT2 à l'aide de l'algorithme XGBoost très efficace. Le modèle proposé détecte avec précision la présence ou l'absence de la maladie sur la base des données d'électrocardiogramme à dérivation unique chez les patients de l'ensemble de données

CARTaGENE. Cela peut constituer un moyen efficace d'effectuer un dépistage du DT2 au niveau de la population, de faciliter la détection et l'intervention précoces de la maladie et d'améliorer les résultats pour les patients.

Acknowledgements

This work would not have been possible without my wonderful supervisors and mentors, Dr. Abhinav Sharma, Dr. Mitchel Benevoy and Dr. Matthias Friedrich. Your collective leadership, intellect, and authenticity made my experience a great one. Thank you to the entire DREAM-CV Lab for the support and guidance, especially Dr. Elite Possik. Conducting this research has truly been an honour and privilege, and I submit this work held up by the love of those around me and those that came before me. To my father, Abdul-Latif, thank you for nurturing my curiosity of the world around me, teaching me the value of education, and always reminding me to live my life with a little bit of poetry. To my eldest brother and one of my greatest mentors, Omar, thank you for walking by my side every step of the way. I cannot recall a moment where I needed you and you were not there, and that is a beautiful gift from someone with your wisdom. To my mother, Khadejah, thank you for reminding me of who I am and our roots every step of the way, and lifting my head up when it hung heavy. To my dearest Nagad, Layla, and Sara, thank you for your sisterhood and for being a light in some of my darkest times – your unwavering support is something I am forever grateful for. To the rest of my loved ones (Susu, Moe, Sandy, Asim, Selena, Kareem, Amira and Medina), your love inspires me every day.

"There is a treasure of knowledge inside every person; an irreplaceable gem that can benefit the world and continuously update and renew itself—only for those who choose to unearth it."

- Albert M. Jabara

Contribution of Authors

The author, Mariam Jabara, confirms sole responsibility for all the contents within this thesis, including study design, model development, analyses, interpretation of results, and manuscript preparation. Revisions were done by Mariam Jabara, Dr. Abhinav Sharma and Dr. Elite Possik.

Abbreviations

AI: Artificial Intelligence

AUROC: Area Under the Receiver Operating Curve

BMI: Body Mass Index

CPU: Central Processing Unit

DL: Deep Learning

ECG: Electrocardiogram

FPG: Fasting Plasma Glucose

GLUT4: Glucose Transporter Type 4

GPU: Graphic Processing Unit

HbA1c: Glycated Hemoglobin A1C

LMIC: Low to Middle Income Country

NPV: Negative Predictive Value

OGGT: Oral Glucose Tolerance Test

PPV: Positive Predictive Value

PCOS: Polycystic Ovarian Syndrome

PSR: Physiological Stress Response

T2DM: Type 2 Diabetes Mellitus

XGBoost: Extreme Gradient Boosting

List of Figures

List of Tables

1 Introduction

1.1 BACKGROUND

Diabetes is one of the largest global health emergencies of the 21st century, with type 2 diabetes mellitus (T2DM) accounting for 90% of all diabetes worldwide (1). In Canada, T2DM presents a significant challenge for our healthcare system and greatly impacts quality of life and longevity for those diagnosed. In 2015, the estimated prevalence of diabetes was 3.4 million and is predicted to rise by 44% to 5 million by 2025 (2).This increase represents a cost of over 15 billion dollars over a 10-year period (3). A key factor in managing the diabetes epidemic is early diagnosis and treatment of the disease. In most cases, T2DM is diagnosed after a clinical event (e.g., a heart attack) or during routine screenings for patients of older age. However, T2DM can emerge much earlier in life and is often unrecognized, with its onset occurring at least 4-7 years before a clinical diagnosis (4). In fact, up to half of those with the disease are currently undiagnosed, and 87.5% of all undiagnosed cases of the disease are in low to middle income countries (LMICs) (1). The ability to screen for and detect the disease in its earliest stages can improve the lives of patients through early intervention and mitigate the burden on our healthcare system by controlling disease progression. Further, accessible screening tools may help to reduce the very high number of undiagnosed individuals in LMICs. Complications of T2DM are widespread across various systems in the body, including but not limited to cerebrovascular disease, retinopathy, heart attack, kidney

damage, and peripheral neuropathy. Consequently, there is a largely unmet clinical need for low-cost, non-invasive methods to screen for T2DM and pre-diabetes. Machine learning (ML), a subset of Artificial Intelligence (AI), is best known for its ability to detect patterns in complex data, making it an ideal method for analyzing patient data to identify undiagnosed T2DM patients. Given that T2DM is a data-rich condition with various outcomes, ML models can be leveraged to learn from previous patient data and aid in early-detection and treatment of the disease (5). While applications of AI in diabetes are popular for management and monitoring of disease progression, AI has yet to be successfully applied to large-scale screening for the disease itself. Electrocardiograms (ECGs) are data-rich and when combined with AI may serve as an inexpensive, scalable, and effective method to screen for T2DM in the general population. In this thesis, we seek to explore an efficient ML model for the screening of T2DM using single-lead ECG data from a standard 12-Lead ECG. We explore the use of time series based, frequency-based, and interval-based features using a highly efficient and scalable ML model for the detection of T2DM.

1.2 THESIS AIMS

1. Review the current digital tools available to screen, predict or diagnose type 2 diabetes using electrocardiograms and machine learning.
2. Develop and validate the use of a machine learning model for predicting diabetes from single-lead electrocardiogram data from a standard 12-Lead ECG.

1.3 RESEARCH QUESTIONS

1. How are ML-based based tools for electrocardiograms used in screening for diabetes and/or pre-diabetes in the general population, and what is the clinical utility of these tools?
2. Can a ML-based tool be used to accurately screen for T2DM using Lead-I from a 12-lead electrocardiogram alone?

2 Literature Review

2.1 Pathophysiology Of T2DM

T2DM is characterized by hyperglycemia and is a result of the body's inability to adequately respond to insulin (1). In comparison to type 1 diabetes, the ability to produce insulin endogenously is partially preserved in T2DM, therefore, most patients are not insulin-dependent (6). Generally, the disease is marked by a combination of 1) deficient insulin secretion by pancreatic islet β-cells, 2) tissue insulin resistance (more commonly), and 3) consequent to points 1 and 2, an inadequate response to insulin secretion (7).

2.2 Glycemic Control

Insulin is a hormone that plays one an important role in glucose homeostasis and metabolism. The synthesis and secretion of insulin are largely controlled by circulating glucose levels. In healthy individuals, glucose stimulated insulin secretion is bi-phasic (8). In the first phase, increased levels of glucose induce the secretion of insulin from β-cells rapidly, typically within one minute. The second phase has a slower onset, and persists for as long as glucose levels are elevated above normal levels (5 mmol/L) within the body (9), (10). Insulin then binds to specific insulin receptors on cell surfaces to increase the storage of glucose by increasing its uptake into fat and muscle cells from the bloodstream. The binding to these receptors activates a complex signal transduction pathway with many endpoints, one of which controls traffic of the transport of glucose

transporter type 4 (GLUT4) receptors from intracellular stores into the plasma membrane. Once glucose enters the cell, it is converted to glucose-6-phosphate, and then can be used either as a source of energy or converted to glycogen for storage (11). Glucose uptake, use and storage occurs in almost all cells within the body, but largely in skeletal, adipose and liver cells (12). When blood glucose levels begin to fall between meals, the glucose stored in the liver as glycogen will be released back into the bloodstream through glycogenolysis, as mediated by the other major hormone in blood glucose homeostasis – glucagon. Glucagon has a profoundly hyperglycemic effect, increasing blood glucose levels within minutes through the breakdown of liver glycogen and gluconeogenesis (synthesis of glucose inside the liver).

2.3 Global Incidence and Prevalence

Currently, there are an estimated 462 million individuals worldwide with T2DM (13) and this number is projected to grow to over 590 million by 2035 (14). Those who are between 40-60 years of age are at highest risk of developing T2DM, however T2DM can present itself in adolescents and children. In fact, T2DM accounts for 45% of newly diagnosed paediatric diabetes in the USA (14). Given that ~ 90% of patients are obese or overweight at T2DM diagnosis, the aetiology of the disease is believed to be linked to risk factors such as sedentary behaviour and poor eating habits (15). Previously, T2DM was thought of only as a disease of the West, induced by high-calorie diets and sedentary lifestyles and in those of older age (14). Recently, however, two thirds of all cases are in

low to middle income countries and the disease has become more common in the paediatric population (16), raising questions about modifiable and non-modifiable risk factors for the disease. Interestingly, in Western populations, up to half of patients with T2D have a BMI >30, and 30-40% have a BMI of 25-30 (where a BMI > 25 is overweight) (17). However, in some Asian populations, approximately half of the patients are not overweight (18). This has since promoted increasing investigation into the interactions between genetics, environmental, and lifestyle factors as they contribute to the development of the disease. Overall, there is an alarming trend of an increase in both prevalence and incidence of the disease over time (Figure 1) (19).

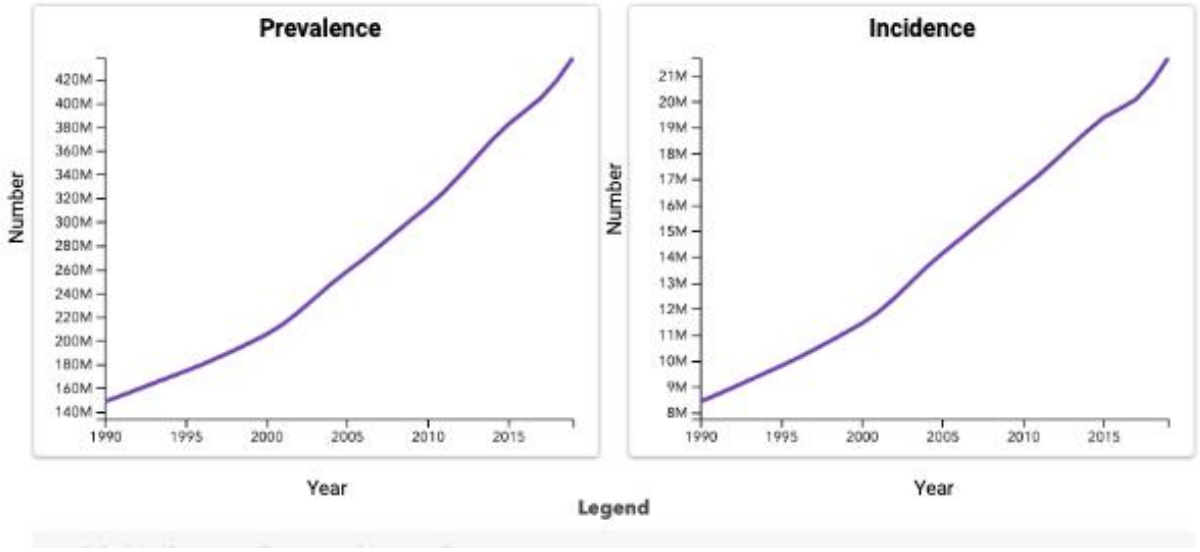

Figure 1. Global prevalence and incidence of T2DM from 1990-2019.

Based on data from the Institute of Health Metrics and Evaluation (20). Created with BioRender.com

Further, it has been found that psychosocial factors such as stressful working conditions, traumatic events, and mental health disorders can increase the risk of developing T2DM. Kelly and Ismail (21) produced a comprehensive literature review of longitudinal studies focusing on the links between psychosocial factors and risk of T2DM. With the growing body of information that T2DM is not the disease we once thought it was, a gap in knowledge exists around additional factors that may be influencing the increasing prevalence and incidence of the disease. In their review, they propose that chronic activation of the physiological stress response (PSR) may increase the risk of developing T2DM. In fact, there is discussion around the role that inflammation plays in T2DM as demonstrated by the elevated levels of inflammatory factors in patients with the disease (22). The authors of this review posit that the inflammation induced by the PSR may contribute to the development of T2DM. In their analyses, they found statistically significant positive correlations between the scores of tools such as the Centre for Epidemiologic Studies Depression Scale (CES-D Scale), Short-Form 36 (SF-36) questionnaire, Zung self-rating depression inventory, burnout, and psychological distress (22).

While the need to understand better the fundamentals behind the aetiology of the disease are critical, there is an equally critical need to identify T2DM patients earlier on and facilitate the well-established, evidence-based strategies to manage the progression of the disease. The combination of an increased understanding in the pathology of the

disease and ability to identify the disease earlier is important in identifying patients at risk, and intervening with novel, evidence-based approaches to reduce risk of developing the disease and slow the progression of it.

2.4 RISK FACTORS

The risk factors for T2DM can be considered as categorized as modifiable and non-modifiable (Tables 1 and 2) (23) (24)**:**

Table 1. Non-modifiable risk factors for T2DM

Risk Factor	Details
Ethnicity	African, Arab, Asian, Hispanic, Indigenous, South Asian
Family History	Increased risk if parent(s) or sibling(s) have diabetes
Age	> 40 years of age
Gestational Diabetes	Increased risk for mother and baby

Table 2. Modifiable risk factors for T2DM

Risk Factor	Details
Diet & Lifestyle	Unhealthy eating habits, high stress, alcohol consumption, smoking, poor sleep hygiene
Obesity	BMI > 30
Other diseases	Including but not limited to: High Blood Pressure, High Cholesterol, Polycystic Ovarian Syndrome (PCOS), Psychiatric Disorders, Sleep Apnea (non-exhaustive).

A significant risk factor for T2DM is obesity, high body mass index (BMI) or high percentage of body fat (specifically around the abdominal region). Higher levels of

adipose (fat) tissue can result in insulin resistance through inflammatory mechanisms such as increased free fatty acid release (7). This is a modifiable risk factor and presents a means for early intervention to prevent the manifestation of T2DM or slow its progression. In fact, nutrition therapy can reduce HbA1c values by 1-2% (2), and lifestyle changes overall have been found to be more effective than pharmacotherapy in reducing disease incidence in some cases (25). Even modest weight loss of 5-10% of initial body weight can substantially improve insulin sensitivity, blood pressure, HDL-C and triglycerides, representing not only an improvement in disease state, but a reduction in risk for complications such as cardiovascular disease (26).

2.5 TYPE 2 DIABETES DIAGNOSTIC GUIDELINES

In Canada, the current diagnostic guidelines for Type 2 Diabetes for adults are as follows (Table 3) (2):

Table 3. Canadian Diagnostic Guidelines for T2DM

Parameter	Threshold
Fasting Plasma Glucose (FPG	≥ 7.0 mmol/L
HbA1c	≥ 6.5%
2-hour plasma glucose (2h-PG) in a 75g Oral Glucose Tolerance Test (OGTT)	≥ 11.1 mmol/L
Random plasma glucose (RPG)	≥ 11.1 mmol/L

Canadian guidelines suggest that upon diagnosis, the following should take place for each patient: identify a personalized target for HbA1c levels, assess their cardiovascular and renal function, educate on diabetes, and healthy behavioural interventions provided. Typically, patients are not started on pharmacotherapy unless their individualized HbA1c target is not met within 3 months of the lifestyle changes. It should be noted that healthy behaviours and weight loss can have a significant impact on a patient's disease state. In fact, patients who start pharmacotherapy may improve their conditions behavioural and weight loss interventions, which can lead to withdrawal of pharmacotherapy and even remission in some cases (27). Additionally, remission can be observed in those who undergo bariatric surgery. In one study, 72% of T2DM patients achieved diabetes remission 2 years after their surgery (28).

There are advantages and disadvantages to each test, but common to many of the tests are disadvantages in cost and convenience (Table 4) (2).

Table 4. Advantages and disadvantages of diagnostic tests for diabetes. From Diabetes Canada Clinical Practice Guidelines (2)

Parameter	Advantages	Disadvantages
FPG	• Established standard • Fast and easy • Single sample • Predicts microvascular complications	• Sample not stable • High day-to-day variability • Inconvenient (fasting) • Reflects glucose homeostasis at a single point in time
2hPG in a 75 g OGTT	• Established standard • Predicts microvascular complications	• Sample not stable • High day-to-day variability • Inconvenient • Unpalatable • Cost
A1C	• Convenient (measure any time of day) • Single sample • Predicts microvascular complications • Better predictor of CVD than FPG or 2hPG in a 75 g OGTT • Low day-to-day variability • Reflects long-term glucose concentration	• Cost • Misleading in various medical conditions (e.g. hemoglobinopathies, iron deficiency, hemolytic anemia, severe hepatic or renal disease) • Altered by ethnicity and aging • Standardized, validated assay required • Not for diagnostic use in children and adolescents† (as the sole diagnostic test), pregnant women as part of routine screening for gestational diabetes‡, those with cystic fibrosis or those with suspected type 1 diabetes
***2hPG*, 2-hour plasma glucose; *A1C*, glycated hemoglobin; *CVD*, cardiovascular disease; *FPG*, fasting plasma glucose; *OGTT*, oral glucose tolerance test.**		
∗ Adapted from Sacks D. A1C versus glucose testing: a comparison (43).		
† See Type 2 Diabetes in Children and Adolescents chapter, p. S247.		
‡ See Diabetes and Pregnancy chapter, p. S255.		

Additionally, all forms of testing are invasive, meaning that they require bloodwork to measure the values for FPG, HbA1c, 2h-PG, and RPG. The ability to screen for T2DM using ECG data would eliminate the need for blood tests that are invasive and costly to use as a screening tool at the population level. By nature, screening tools should be cost-effective, scalable, and easy to acquire. The cost of a glucose blood test in Canada was $19.15 (29) in 2015, a price that includes both the cost for the blood draw, and the single test for glucose levels ($15.62, and $3.53, respectively). To obtain HbA1c values, a test would cost $23.82, with the HbA1c test costing $12.69. In comparison, the cost of an electrocardiogram $11.05 (30) but this value includes interpretation of the results. In a scenario where an ECG is being used as a screening tool for T2DM, the cost for interpretation may be reduced as it would be automated by the machine learning algorithm. In addition to proposed cost-effectiveness, the ability to performing screening in rural areas may be increased using signal-based screening methods. For those living in rural parts of Canada, these costs are increased as patients may need to be flown out for routine labs due to access issues to healthcare in their region. Therefore, screening using ECG data may make accurate, population level screening for T2DM a reality in both urban and rural region, with great potential for scalability and cost-savings.

Most forms of screening involve the calculation of risk scores. In Canada, CANRISK was developed by the Public Health Agency and is used to evaluate risk for pre-diabetes and type 2 diabetes in those aged 40 years or older (31). The values of the

risk score are related to the 10-year risk for developing T2DM, with scores between 0-32 suggesting a 1-17% risk, scores between 33-42 suggesting a 33% risk, and scores between 43-87 suggesting a 50% risk. The calculator takes inputs such as age, sex, BMI, waist circumference, family history of diabetes, history of high blood sugar, history of hypertension, daily activity levels, consumption of fruits and vegetables, gestational diabetes risk (for females), ethnicity, and level of education. In a study that validated the CANRISK model in a multi-ethnic population, the area under the receiver operating cure (AUROC) was found to be 0.75 (95% CI: 0.73-0.78) (32). It should be noted that the CANRISK tool was not developed or validated in those younger than 40 years of age.

2.6 SCREENING: GLYCATED HEMOGLOBIN (HBA1C) AND FASTING PLASMA GLUCOSE (FPG)

In Canada, screening for T2DM is to be conducted through measurement of fasting plasma glucose and/or glycated hemoglobin every 3 years in individuals 40 years of age or older or those who score in the "high risk" category on the CANRISK calculator (2). HbA1c reflects the average plasma glucose over the previous 8-12 weeks (33), and the FPG test reflects the body's ability to regulate glucose levels via insulin and glucagon – where a high value suggests some form of insulin resistance. The HbA1c test can both be used for screening and confirmatory testing for diagnosis, and additional testing may be required to rule in the disease if the patient's results are at the borderline (6.5%) for the test.

In a 2022 retrospective population-based study of over 1 million adults, it was found that adherence to screening guidelines was suboptimal (34). Furthermore, the cut-off age for screening may need to be re-evaluated as the early onset type 2 diabetes is becoming more prevalent (35). In fact, early onset T2DM can increase the risk of complications such as myocardial infarction by up to 14 times compared to control subjects (36).

2.7 T2DM AND THE CARDIOVASCULAR SYSTEM

Based on the work of Harris et. al, the onset of non-insulin dependent diabetes mellitus (NIDDM) may occur up to 4-7 years before clinical diagnosis, based on a study that examined retinopathy images of over 5000 patients from the United States and Australia (4). Additionally, early signs of cardiac autonomic neuropathy may be identifiable through subtle ECG alterations (37). At more advanced stages of T2DM progression, cardiovascular disease may develop as a complication. The combination of insulin resistance, hyperinsulinemia and hyperglycemia triggers a signalling transduction pathway involving multiple cellular and molecular pathophysiological factors that ultimately lead to increased levels of inflammation and endothelial dysfunction. This presents an ideal environment for the development of atherosclerosis and atherosclerotic cardiovascular disease. Therefore, at all stages of T2DM progression, there are both major and minor effects on the cardiovascular system, both of which may be identifiable through techniques such as machine learning.

Machine learning presents an opportunity to develop novel approaches to screening for T2DM and other chronic diseases. As mentioned, T2DM is a data rich condition and machine learning can be used to detect patterns in electrocardiogram data that would be impossible to extract using other statistical methods or the naked eye. Therefore, the combination of the effects T2DM on the cardiovascular system, increased availability of digital data from mobile and wearable devices, and the need for more effective and scalable screening strategies presents the ideal opportunity to apply machine learning for the early detection of T2DM.

2.8 Related Work

There are several approaches that have been developed to detect, monitor, or diagnose type 2 diabetes from electrocardiographic data; however, these are largely centered around deep learning (DL) models, require a full 12-Lead ECG, or require a combination of bio-signal data (38), (39), (40), (41).

Table 5. Overview of literature review findings

Authors	*Model Architecture*	*AUROC*	*Accuracy*	*Sensitivity (%)*	*Specificity (%)*
Lin et. al	Deep Learning	82.55	N/A	N/A	N/A
Dave et. al	Random Forests	86.5	N/A	79	79
Nguyen et. al	Artificial Neural Network	N/A	N/A	70.59	65.38
Gupta et. al	Intrinsic Time Scale Decomposition	N/A	86.9	84.56	90.6

In the work of Lin et al (38), the electrocardiogram was used to estimate the value of HbA1c for patients using DL. Additionally, they explored the use of XGBoost and elastic net – the former being a non-deep learning ML model, and the latter being a penalized linear regression model. Their DL model was able to achieve an AUROC of 0.8255 for DM screening on their testing cohort. However, their models were not specific to type 2 diabetes mellitus, and included patients with type 1, type 2, gestational or specific types of diabetes due to other causes. The architecture of their best DL model involved an "attention mechanism", which borrows its name from the attention mechanism in human visual processing. For example, when we read a book, most of our focus (attention) is on the word we are currently reading and trying to process, and not the remainder of the text on the page. Similarly, we would like the DL model to focus on

the most relevant parts of the input while making a prediction. The attention mechanism or "module" is an additional network layer on top of an existing DL model that is used to boost the accuracy of predictions (42). In their best performing model, a positive correlation between actual HbA1c score and ECG derived HbA1c (ECG-HbA1c) was found, with r=0.577 (95% CI: 0.531-0.582). In subgroup analyses, it was found that patients with higher ECG-HbA1c had more risk factors for DM progression. This alludes to the potential for machine learning based approaches to provide more granular information in comparison to HbA1c. In conclusion, their work demonstrates the results of a novel biomarker estimating HbA1c in various forms of diabetes and compares a DL and non-DL based approach with promising results. However, due to the inclusion of all types of diabetes in this study, the model cannot provide insight into the type of diabetes a patient has – an important characteristic for treating and managing disease progression.

The work of Dave et al. (43) used a combination of electrocardiogram and accelerometer data to predict hypoglycemia and hyperglycemia with data acquired from the Medtronic Sephyr Biopatch, which collects both types of data in fives participants followed for 14 days. They explored the use of both classification models, and regression models. Their classification approach was based on random forests, while their regression approach was based on quantile regression forests. In the classification models, input features from the patients would be used to classify them either as hyper or hypoglycemic. In this study, 70 mg/dL was used as threshold for hypoglycemia and

180 mg/dL was used as threshold to define hyperglycemia. For the regression models, the actual values for glucose levels were predicted. Although there are limitations in the sample size for this study, the authors bolstered the number of training samples by separating each set of data for each patient into five equal partitions – one to be held out for testing, and one for training. They employed a fivefold cross validation technique in this manner to ensure that training and test sets were from non-overlapping time windows. They extracted a total of 45 ECG features and 20 accelerometer features per patient. They built two classifiers – one to detect hyperglycemic events, and one to detect hypoglycemic events. They reported a sensitivity and specificity for detecting hyperglycemia of 79%, and a best AUROC of 86.5. The classifier for predicting hyperglycemia performed better than the classifier for predicting hypoglycemia. This may suggest that the detection of hyperglycemia is more feasible using ML models, and therefore a more suitable target for screening purposes in the context of T2DM. Additionally, they explored the use of ECG features alone, and then a "fusion" model which combined ECG and accelerometer features. In all cases, the fusion model outperformed the ECG-only models. As accelerometer data is easily accessible through wearables and even smartphones, the use of their features may be important in identifying glycemic events and can be used to boost the performance of algorithms where the data is readily available. Overall, this work answers questions around the use of multi-modal sensor data for the detection of a T2DM related parameter: glycemic

control, through regression and classification-based approaches, but is limited by the small sample size (n=5) and by the number of hyper/hypoglycemic episodes in the participants. Additionally, all 5 participants were healthy, lending questions to how such a model would perform amongst those with any form of diabetes.

Nguyen et al. developed a neural network approach for non-invasive detection of hyperglycemia using ECG signals in Type 1 Diabetic patients (40). Although the pathophysiology and complications of type 1 diabetes differs from those of T2DM, it is still important to consider models developed across the diabetes spectrum to better understand how physiological signals related to glycemic control. In their work, they developed a feed-forward multi-layer neural network model to detect the presence of hyperglycemic events. Ten type 1 diabetic patients. For each patient, a 30-minute segment of ECG data was recorded, while blood glucose samples were collected at the same time at regular intervals. For each patient, sixteen ECG features were derived, five of them being standard intervals that are easily observed from the ECG data (and are typically automatically generated by most 12-Lead ECG machines), and the remaining 11 features being a combination of time series and frequency domain features. For their results, the test set lead to 70.59% sensitivity and 65.38% specificity in identifying hyperglycemic events. Further, they compared their neural network model to two other models commonly used for classification: linear discriminant analysis and K-nearest neighbours. Their findings suggested that their neural network model was superior to the other

methods for hyperglycemia classification. This work demonstrates a unique approach to combining interval-based, time-based, and frequency-based features using a neural network-based model to detect hyperglycemia in patients with Type 1 Diabetes. This work is limited by its sample size (n=10) and the monitoring of blood glucose levels during an overnight period only. Future works should explore the use of this model in patients with Type 1 Diabetes for longer periods of time, as well as it's applicability to patients with other forms of the disease (T2DM, gestational diabetes, etc.).

Gupta et al. proposed a framework for automating screening of type 2 diabetic patients using ECG signals (41). They used single-lead electrocardiogram data and trained a decision tree classier on data from 35 patients with diabetes and 51 control patients. However, to boost the number of training samples, each 20–50-minute ECG reading was segmented into 5-second fragments and labelled as diabetic or normal, and it was not clear if patients from the test and train sets were kept separate. To extract relevant features for the 5-second fragments, intrinsic time scale decomposition (ITD) was used to decompose the signal into its rotational components. They did not report AUROC, but reported an accuracy of 86.90%, specificity of 90.60%, and positive predictive value of 84.56%. Overall, their work explores the use of single-lead ECG data for screening of T2DM through intrinsic time scale decomposition and provides promising results. However, because it was not clear whether splitting of the data for training and testing was performed at the ECG fragment level, or patient level, validation

with an independent test set should be conducted to explore the generalizability of the model and confirm the observed results.

Despite the strong trend of using deep learning and neural network-based approaches to detect T2DM and its associated parameters, there is a significant gap in the literature around scalable tools for T2DM detection. Further, there is no consensus on which features of the ECG are most appropriate to use, as what degree of explain ability is required for these models to be implemented in clinical settings. Typically, in deep learning, the neural network model can be thought of as a feature extractor where features are being extracted from the time series to predict the target variable – presence or absence of T2DM. However, these features may not be medically relevant or explainable in the context of the anatomy or physiology of the cardiovascular system and T2DM. This problem becomes increasingly relevant as we consider the fact that complex machine learning models are prone to overfitting, which is a term that describes models which are too closely fit to the training data, and therefore do not perform well on unseen data. Deep learning architectures are inherently more complex, and therefore the risk for overfitting is greater. When selecting a model architecture to solve a problem with machine learning, more complexity does not necessarily lead to better outcomes. Similarly, an overly simplified model cannot capture the trends in the data to make accurate predictions. The design of an effective model is a delicate balance between model complexity and appropriate feature engineering. Feature engineering involves

both the selection of features already present in the dataset (feature selection), and new features that are derived from already existing features (feature extraction). Evidently, the design of an accurate model for the early detection of T2DM should consider model complexity, scalability, explosibility, and feature engineering.

The scalability of a model can be assessed in the development of the model, where one would consider how the model scales with increasing amounts of training data. Considerations at this stage are largely related to training time and the computational power required (how many machines are needed, is distributed learning necessary, etc.). Additionally, post-training considerations are of great importance; how long does the model take to produce predictions, and can the model be deployed on smaller devices such a mobile phones or smart watches? It is imperative that research on the use of ML for early detection of T2DM is forward-thinking, and considers the implications of model choice on scalability, accessibility, and amount of training data required. While neural network-based models may produce better predictions in some scenarios, when they are deployed at scale, the performance may decrease due to downsizing of the model architecture, computational efficiency, or limitations in access to training data. If similar results can be achieved with lighter-weight models such as XGBoost, this would result in significant benefits at later stages in terms of cost, infrastructure, and efficiency. Therefore, in this work, we seek to address this issue by developing a non-deep learning approach to predict T2DM using XGBoost as model of choice. We compared the use of

interval-based, time-series based, and frequency-based features as input to the model to examine their utility to predict T2DM.

When designing a study on the use of electrocardiogram data for T2DM detection, the question of how to partition the data becomes ever relevant. Most datasets contain ECG readings of 5-10 seconds in length for each patient; however, many previous studies have opted to partition each ECG reading into smaller fragments (with each fragment representing one cardiac cycle or one heartbeat). This is an effective way to improve the size of the training dataset, with an increase up to 10-fold. However, if individual heartbeats are being used as single samples, there should be no overlap of heartbeats from a single patient in both the training and testing datasets. The goal of the aforementioned models, as well as those being developed in this thesis, is to predict the presence of T2DM from unseen data based on the previously seen data used to train the model. In practice, if there is overlap of samples from the same patient in training and testing, then theoretically the model may be learning representations of T2DM specific to one patient, and then be tested on heartbeats from the very same patient. This may lead to unusually high results for AUROC, specificity and sensitivity. To validate such models, testing on an additional dataset with no patient overlap with the dataset used for training should be conducted. Therefore, results from studies where it is unclear whether the patients were separated in the training and testing sets should be interpreted with this in mind.

3 Methods

3.1 Overview

To evaluate the use of machine learning for the screening of T2DM, we developed a model that uses Extreme Gradient Boosting (XGBoost) in conjunction with various data pre-processing/feature extraction methods (Figure 2). We built a binary classifier using XGBoost that takes Lead-I of a 12-Lead as input, and outputs a label for patients from the following classes: (0) diabetes negative or (1) diabetes positive. All models were trained on a MacBook Pro (2018) with 2.3 Ghz Quad-Core Intel Core i5 processor and 8 GB memory. All code was written in Python 3.9 in Microsoft Visual Studio using Jupyter Notebooks. Pre-processing was done using Numpy and Pandas packages.

We sought to develop a novel approach to applying XGBoost to time series data, as the model is typically used for complex datasets with tabular data and columns that represent *time-independent* features. However, as ECG data is inherently *time dependent*, we extracted time-independent features using various pre-processing and dimensionality reduction techniques to produce inputs to the XGBoost model. We evaluated feature extraction and pre-processing methods across three categories: 1) time domain analysis, 2) frequency domain analysis, and 3) automatically extracted ECG features. For all experiments, we tested the model with ECG features alone, as well as

ECG features in combination with the following patient metadata: age, sex, and body mass index (BMI).

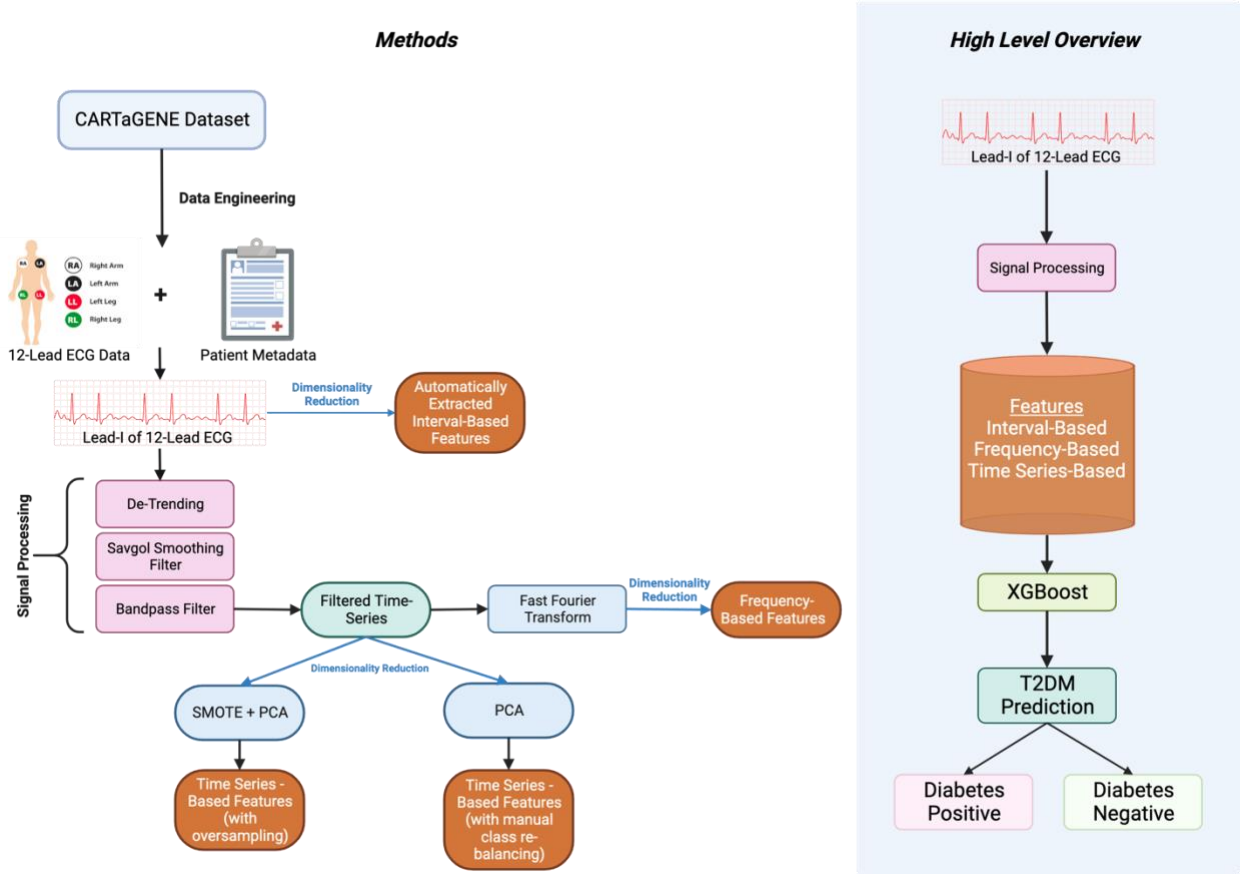

Figure 2. Methodology overview. Created with BioRender.com

3.2 DATASET

For all experiments, we trained and validated the model using data from the CARTaGENE study from CHU St. Justine (REB# 2017-3010). CARTaGENE is a public research platform that aims to accelerate health research. The platform contains both biological samples and data on the health and lifestyle of 43,000 Quebec men and women between the ages of 40 and 69 at recruitment. The data was collected between 2009-2010,

and includes electrocardiogram data for N=7463 patients, and of these patients N=635 have T2DM (Table 6). ECGs were acquired using the GE Cardiosoft electrocardiogram machine and software and sampled at 500Hz. In this dataset, T2DM is defined as self-reported diagnosis OR HbA1c value ≥ 6.5% (in line with current Canadian guidelines for T2DM diagnosis).

Table 6. Demographics for CARTaGENE dataset

Demographic	**Variable**	**Overall**
Age, years	Median (IQR)	53 (49-61)
	Mean ± SD	55.2 ± 7.6
Gender	Female	379
	Male	284
Body Mass Index	Median (IQR)	28.3 [25.2, 32.7]
	Mean ± SD	29.3 ± 5.8
Race	Non-Black	654
	Black	9
Avg. Systolic Blood Pressure	Median (IQR)	124 [113, 133]
	Mean ± SD	124 ± 15
Avg. Diastolic Blood Pressure	Median (IQR)	74 [66, 80]
	Mean ± SD	74 ± 11
Glycated Hemoglobin (HbA1c)	Median (IQR)	5.9 [5.5, 6.7]
	Mean ± SD	6.3 ± 1.39

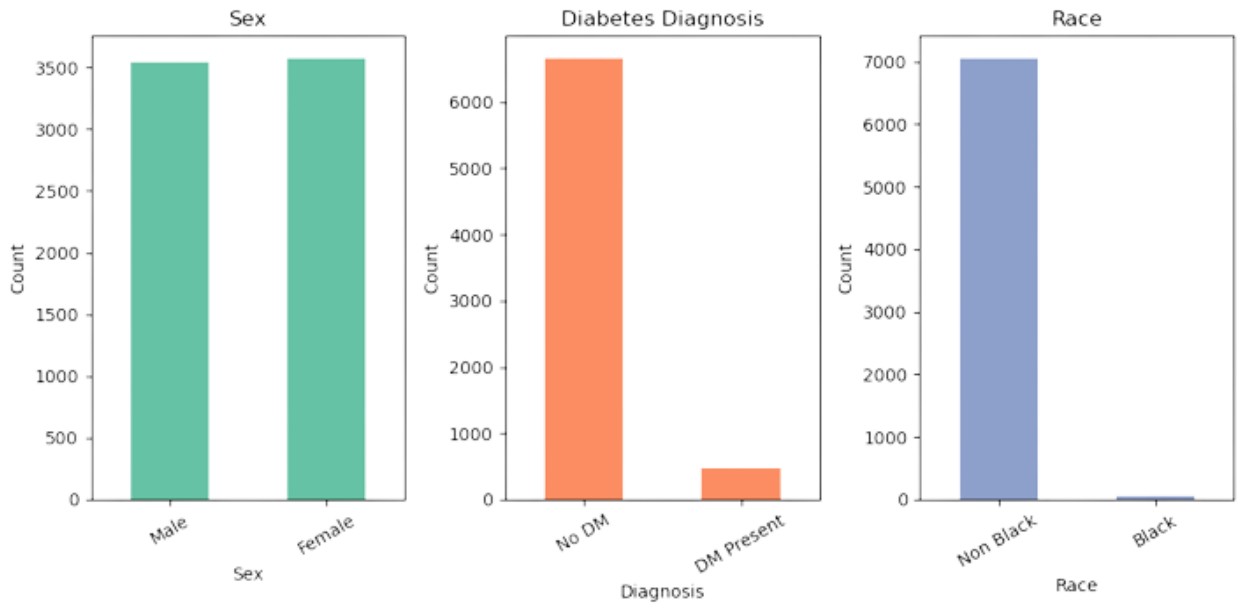

Figure 3. Visualized demographics for CARTaGENE dataset.

3.3 Model Evaluation

For each experiment, we calculated the area under the receiver operating curve (AUROC) sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and F1 Score for each classifier. Sensitivity and PPV are often referred to in the Computer Science literature as recall and precision, respectively. The AUROC and F1 metrics are also more common in the field of machine learning, both describing the discriminative ability of the model. The AUROC describes a model's classification abilities at different thresholds for classification by plotting the false positive rate against the true positive rate (or sensitivity against 1-specifictiy). The area under the curve represents the degree of separability between classes in terms of how the model classifies samples. For example, if a model can always correctly classify T2DM patients as having

the disease, and those without the disease as not having it, then the degree of separability between classes is perfect, which corresponds to an AUROC of 1. A binary classifier that *randomly* classifies patients as having the disease or not having the disease would have an AUROC of 0.5, meaning that only 50% of the time patients are correctly classified. Therefore, values of AUROC generally fall between 0.5-1.0. The F1 score is formally defined as the harmonic mean between the PPV (precision) and sensitivity (recall). A harmonic mean calculation will provide a lower score when either value is very low, or there is a large difference between the two values. This calculation is important as when increasing PPV, sensitivity may suffer and vice versa. Therefore, a high F1 score would indicate a well-balanced classifier, where the model has suitable performance when it produces a positive prediction (in other words, the PPV is adequate), and can adequately classify positive patients overall (sensitivity).

3.4 MODEL ARCHITECTURE: XGBOOST

For all experiments, we trained an XGBoost model and performed a grid-search for hyperparameter optimization. XGBoost is an efficient implementation of Gradient Boosting Trees: a machine learning algorithm that combines sequentially connected 'learners' to produce predictions (44). Compared to deep learning, an architecture based on deeply layered neural networks, XGBoost often requires less data to train to achieve similar results. Given that data acquisition to train models can be difficult in clinical settings, we wanted to explore an approach to the T2DM screening problem that did not

require DL and was easily scalable and efficient for both training and inference. The sequential training of individual learners is described by the term 'boosting' in gradient boosting trees and is built upon the concept of additive modelling: a complex function to describe the data is derived from the combination of more simple functions. Gradient boosted models take the weighted sum of the sequentially trained learners to produce a final function that describes the dataset in a way that minimizes the **loss function**. In building supervised learning models, the overall goal is to find a model that minimizes loss (45), a value that indicates how poorly the model performed at making a prediction on a single example. There are various loss functions that can be chosen, but all have the goal of evaluating the model's ability to make correct predictions. High loss would indicate poor predictions, and *vice versa*.

Before discussing the algorithm in detail, the concept of *bias-variance trade-off* must be understood. Bias can be understood as how well the model matches the dataset – high bias meaning the model does not match the dataset well, and low bias meaning it closely matches the dataset. High bias models fail to capture trends in the data, may be overly simplified (underfitting to the training data) and have high error rates. Variance can be defined as the changes in model performance when training on different subsets of the training data. Also, models with high variance tend to perform well on training data, but poorly on test data; in other words, these models do not generalize well on unseen data. Therefore, if we train the model on different subsets of the data, and the model's

performance varies greatly between these subsets, we understand the model has high variance and is likely overly complex (overfitting to the training data). Therefore, there is always a trade-off between bias and variance: a model cannot be both overly complex *and* overly simplified. Similarly, it cannot both fail to capture trends in the training data and capture these trends almost too closely in a way that does not generalize outside the training data.

The algorithm (46) consists of three steps (Figure 3), which can be easily described by following an example, such as the model we have built in this work. The model is used to predict the presence of T2DM from ECG data, with the target variable (what we are trying to predict) being T2DM and the input variable being patient ECG data.

In step one, we initialize a model with a constant value. Initially, we will assume a constant probability across all patients that they belong to the diabetes class. This value will serve as a starting point for building the first tree, which is called a *weak learner*. Weak learners generally perform poorly on their prediction task, but the mistakes of weak learners are the building blocks for subsequent trees. In other words, the mistakes made be the initial trees or *weak learners* are corrected for each subsequent tree, and we end up with a set of trees that makes accurate predictions.

In the second step, we calculate the *residuals*, or the difference between the actual value and predicted value. After the initialization step, we assumed a constant value for the probability that a sample belongs to the diabetes class. For each patient, there is a

ground truth which indicates whether they belong to the positive class (diabetes) or the negative class (no diabetes), which is often denoted by a 1 or 0, respectively. To calculate the residuals, probability that the patients belong to the positive class are subtracted from the actual value (either 0 or 1).

In the third step, we fit a new model on what we failed to correctly classify in the previous model -- in essence, fitting a model on the mistakes of the previous model (47). The final model is a combination of the models that were built sequentially with the previously mentioned algorithm, and the larger "meta" tree is one that has greatly minimized the residuals (errors) (Figure 4).

Input: training set $\{(x_i, y_i)\}_{i=1}^n$, a differentiable loss function $L(y, F(x))$, number of iterations M.

Algorithm:

1. Initialize model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For m = 1 to M:
    1. Compute so-called *pseudo-residuals*:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

    2. Fit a base learner (e.g. tree) $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^n$.
    3. Compute multiplier γ_m by solving the following one-dimensional optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$$

    4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

Figure 4. Gradient boosting algorithm (46)

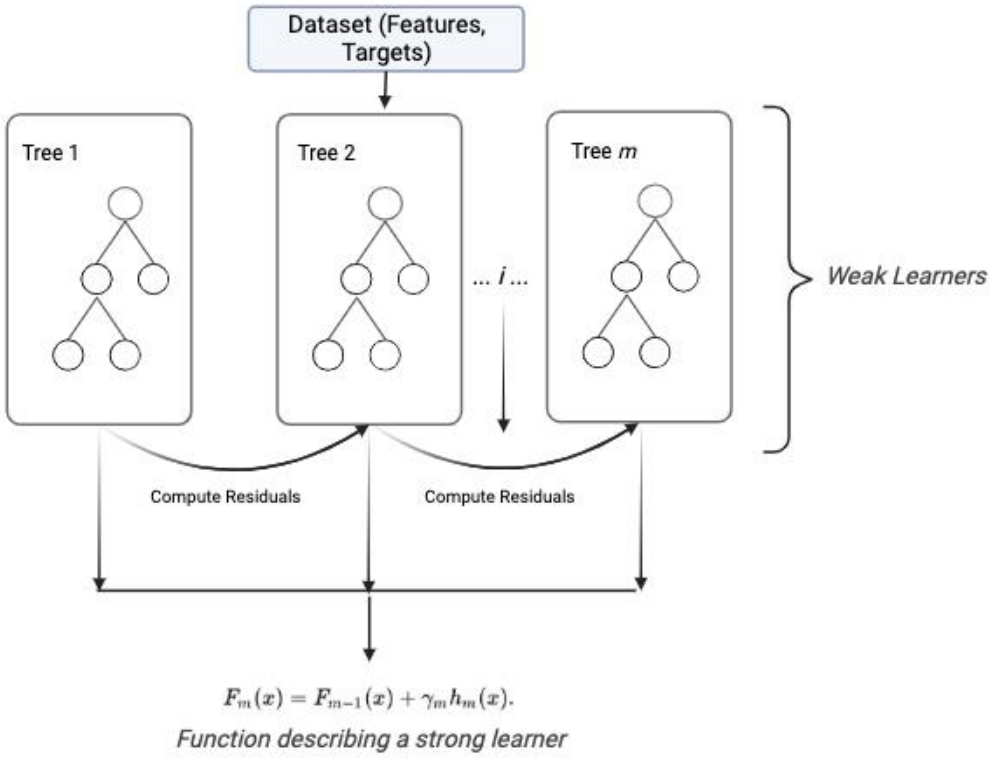

Figure 5. Extreme Gradient Boosting (XGBoost) algorithm overview. Created with BioRender.com

3.5 DIMENSIONALITY REDUCTION

A critical step in the use of XGBoost is dimensionality reduction: transforming data into a lower dimension, while still preserving the information present in the original representation (48). If we include all the data points or 'features' that we have for each patient from each ECG, we risk the model overfitting. Overfitting is a term used to describe a model whose prediction function is too closely fit to the training dataset, and as a result, performs poorly on unseen data (49). Recall that when building a machine learning model, fundamentally, the model is trained on a set of data (training set) and will be tested against unseen data to evaluate its performance. The purpose of the training

data is to serve as a means for the model to understand the relationship between the target variable (what you are predicting) and the input variable(s) (the features you are using the make the prediction). Overfitting is the result of a having an overly complex model, and can be mitigated through proper hyperparameter tuning, in other words, running experiments to choose the most appropriate parameters for the model. The ideal model produces a prediction function which adequately represents the training data, but still generalizes well – in other words, performs well on new examples. The opposite of overfitting is underfitting, and this occurs when the model is overly simple and fails to capture the trends in the training data related to the target variable (50). Models that are underfit typically perform poorly on both training and testing data. This can also be mitigated through adequate hyperparameter tuning, but also may require the selection of a more complex model.

3.6 PRE-PROCESSING

For all experiments, ECGs were pre-processed in the same manner. First, the ECG signal is de-trended, in other words, the mean is removed from the data. This allows us to view sub-trends in the data or unveil cyclical patterns, which in this case, may or may not be related to the presence of T2DM. Next, we apply the Savitzky-Golay (Savgol) filter to smooth the signal, which will remove high frequency and increase the precision of the data without distorting the signal's overall shape (Figure 6) (51). This filter is applied to a signal with N points, and uses a window filter w, that slides across the signal. For each

window segment, the filter will fit a polynomial function of order *o*. Both the window size and polynomial order can be adjusted for more granularity in the control over the level of smoothing and de-noising (51). Finally, we apply a bandpass filter that only allows frequencies between 1-25Hz to pass through, to remove very low and high frequencies which may be attributed to noise while the 12-Lead ECG was being obtained.

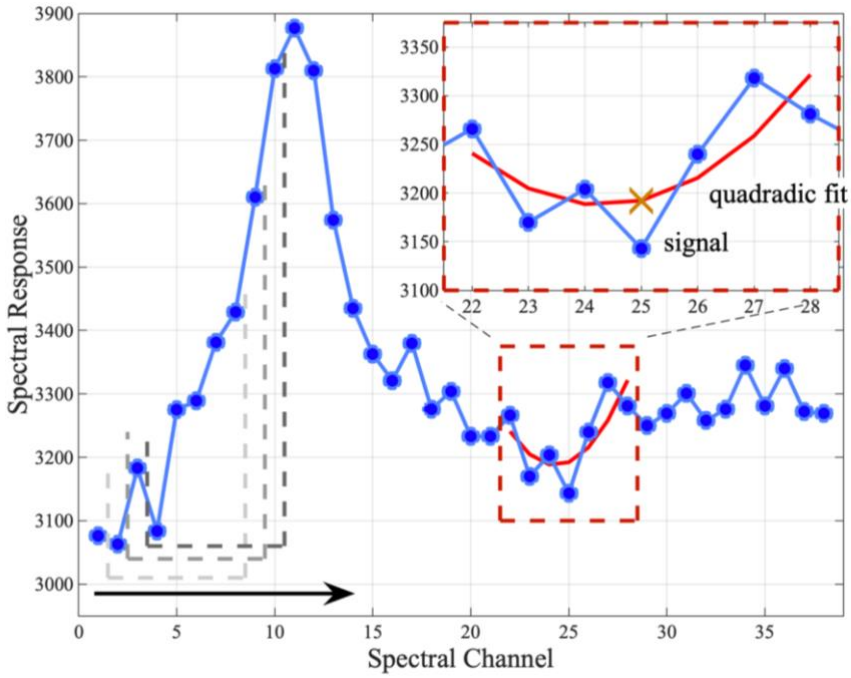

Figure 6. Savgol filter example. The Savgol filter is used to fit a quadratic function incrementally over the entirety of the signal, reducing noise and providing a smoothed version of the signal. (51)

3.7 Principal Component Analysis

Principal component analysis (PCA) is one of the most common methods for dimensionality reduction. PCA is a statistical method that reduces the number of dimensions of the data while preserving the maximum amount of variance (information) (52). Principal components are produced from the input data, with each component

accounting for a specific amount of variance in the data. The number of principal components produced can be controlled by either specifying the number of components desired explicitly, or by indicating a percentage of the total variance in the original data that you wish to maintain. For the purposes these experiments, we used thresholds of to maintain 80% and 90% of original variance in the data.

3.8 Time Series Analysis

For the time series analysis experiments, we sought to perform feature extraction from the time series data directly. To extract features, we used principal component analysis (PCA) as a form of dimensionality reduction, and effectively, as a feature extractor. PCA is one of the most common forms of dimensionality reduction, and is by definition a statistical method that reduces data into "principal components" which explain a certain amount of variance within the data. The amount of variance explained, and in turn, the number of principal components produced by the method can be set manually – either by specifying the number of components explicitly, or by specifying the amount of variance you wish to maintain (from which, a set number of components will be derived). In the experiments using PCA, we specified maintained variances of 80% and 90%, which produced 162 and 236 principal components, respectively.

3.9 Frequency Domain Analysis

For the frequency domain analysis experiments, we employed the Fast Fourier Transform (FFT) to transform the time series data from the time to the frequency domain.

The FFT decomposes a signal into its unique frequency components. Therefore, it provides information on the frequency trends in the data, rather than time dependent trends. In these set of experiments, we are using the FFT as a feature extractor.

3.10 AUTOMATIC INTERVAL-BASED ELECTROCARDIOGRAM FEATURES

The electrocardiogram machine will automatically calculate the following 14 features upon recording a patient's heart activity: PP-interval, PQ-interval, P-axis, P-duration, P-offset, P-onset, QRS-duration, QRS-number, QTC-interval, QT-interval, Q-offset, Q-onset, RR-interval, and R-axis (Table 7).

Table 7.Electrocardiogram interval-based features & descriptions

ECG Feature	Description
PP-Interval	Represents the time between successive P waves on the ECG. It measures the duration of the atrial depolarization and reflects the regularity of atrial rhythm.
PQ-Interval	Measures the time from the beginning of the P wave to the start of the QRS complex. It represents the conduction time from the atria to the ventricles and includes the atrioventricular (AV) node delay. Also known as PR interval.
P-Axis	Refers to the overall direction of the atrial electrical activity. It is determined by analyzing the P wave in multiple leads (53).
P-Duration	Measures the time duration of the P wave.
P-Offset	Represents the endpoint of the P wave, marking the completion of atrial depolarization.
P-Onset	Represents the starting point of the P wave, indicating the beginning of atrial depolarization.
QRS-Duration	Measures the time from the beginning to the end of the QRS complex. It reflects the time taken for ventricular depolarization (contraction of the ventricles) to pump blood out of the heart.
QRS-Number	Refers to the number of QRS complexes observed on the ECG during a specific period.
QT-Interval	Measures the time from the beginning of the QRS complex to the end of the T wave. It represents the total duration of ventricular depolarization and repolarization (54).
QTC-Interval	Represents the heart-rate adjusted QT interval Provides a more accurate assessment of ventricular repolarization due to the variation in QT interval as it relates to heart rate (54).
Q-Onset	Represents the beginning of the Q-wave.
Q-Offset	Represents the end of the Q-wave.
RR-Interval	Measures the time between successive R waves on the ECG. It represents the duration of one cardiac cycle and is used to assess heart rate and rhythm regularity.
R-Axis	Refers to the overall direction of the ventricular electrical activity. It is determined by analyzing the QRS complex in multiple leads

3.11 CLASS IMBALANCE

Given that the data used in this work is highly imbalanced (large number of diabetes negative patients), we must employ a technique to rebalance the classes. If we do not rebalance the classes, the model may simply learn to predict the majority class more often due to its oversampling. We experimented with three methods of correcting class imbalance: 1) Synthetic Minority Oversampling Technique (SMOTE), 2) Manually down sampling the majority class, and 3) Tuning of the parameter which scales the gradient for the positive class (i.e.. punishes the model more strongly for mistakes made against the positive class).

Synthetic Minority Oversampling Technique (SMOTE):

SMOTE is an oversampling technique in which the minority class is over-sampled by creating "synthetic" examples (55). The oversampling is achieved by taking each minority class sample and all its features and introducing new examples along the line segments joining all the *k*-nearest neighbours, with the value of *k* depends on the amount of oversampling needed (the larger the oversampling required, the nearest neighbours are used). The proposed approach involves first computing the difference between the feature vector of the sample under consideration and its nearest neighbor. Subsequently, this difference is multiplied by a random number between 0 and 1. The resulting value is then added to the original feature vector, effectively creating a random point along the line segment connecting the two initial features (55).

For the purposes herein, each time point from the ECG will be considered a time-dependent feature. Therefore, SMOTE will be generating artificial ECGs based on the distribution of the time points for neighbouring samples. However, it is important to understand that SMOTE will interpret each *time-point* as an independent feature. In other words, a new time point will be generated based on values for the same time point in other samples. Additionally, SMOTE will be oversampling based only on the minority class, in this case, the diabetes positive class. This technique was applied to the time series analysis experiments *before and after* applying PCA to examine the effects on synthetic data on raw times series data and its associated principal components, respectfully.

Manual Down-Sampling of the Minority Class:

When manually down-sampling the minority class, the number of negative samples is reduced so that the number of positive samples is equal to the number of negative samples. As a result, the training dataset consisted of N=930 samples with 465 patients in each diagnosis category (diabetes negative and positive). The test set consisted of N=340 samples, with 170 patients in each diagnostic category.

3.12 PARAMETER TUNING:

We selected a Grid Search with Cross-Validation (Grid Search CV) as the method for optimizing the performance of a model. The cross-validation segment of this approach will be further discussed in the following section. The Grid Search is an exhaustive search approach, testing all possible combinations of a specified set of parameters for a model.

The following describes the hyperparameters used and their associated descriptions (Table 8) (56).

Table 8. Hyperparameters of XGBoost and their associated descriptions

Hyperparameter	*Description*	*Values used*
η (Eta)	Learning rate	[0.05, 0.15, 0.2, 0.3]
max_depth	Maximum tree depth	[1, 3, 6, 9]
min_child_weight	Minimum weight for each child in of the tree	[1, 2, 3, 4, 5]
γ (gamma)	Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be.	[0.5, 1, 1.5, 2, 5, 10, 50, 100, 150, 200]
Scale_pos_weight	Amount to scale the gradient for the positive class. Scales errors made by the model during training on the positive class and encourages the model to over-correct them.	[25, 50, 75, 100]

3.13 TRAIN/VALIDATION/TEST SPLITS

The original dataset was split into training and test sets that represented ~5% of the total data. The decision to reserve a smaller test set was made as we had a limited number of positive patients in the dataset to begin with. Approximately 9.2% of patients were T2DM positive in the dataset for the time series and FFT experiments, and 7.0% of patients in the dataset for the automatic ECG features experiment. The cohorts for these

two classes of experiment were different due missing data for automatic ECG features for 356 patients. Most importantly, there was no patient overlap between any of the training and validation data and testing data. This is to ensure the model is learning representations on an independent set of patients, and the discriminative ability is being evaluated on another independent set.

The training set was further divided into smaller training and validation steps for model optimization and hyperparameter tuning. To better evaluate the model's performance and generalizability, we employed a technique called N-Fold cross-validation during the hyperparameter tuning step to expose as many samples from the training data as possible during the hyperparameter exploration (57). The dataset is divided into N subsets or "folds" of approximately equal size, in this case, N=3 folds. First, the dataset is randomly partitioned into the specified number folds (typically 3 or 5). Next, the model is trained and evaluated N times. In each iteration, a different subset of the data is reserved for validation, while the remaining N-1 folds are used to form the training set (Figure 7). The model is trained on the training set, and the performance of the model using the hyperparameters in question is evaluated on the validation set. This is done exhaustively for all possible combinations of hyperparameters as specified by the hyperparameter grid (Table 8). The best model is chosen based on the value of the specific performance metric at each iteration. Models are typically optimized for AUROC, precision, recall, or F1 score. The model is then retrained on the entirety of the training

data using the combination of hyperparameters with the best performance during N-Fold cross validation.

The purpose of this technique is to expose the model to as many subsets of the training data as possible, and to prevent overfitting. When we shuffle the training and validation step at each iteration, we are increasing the likelihood that the model will not learn representation that are overly specific to the training data, thereby increasing the likelihood of producing a model that performs well on unseen data. We reserved a held-out test set, completely independent of the training and validation sets, which are used to evaluate the final performance of the model using the best hyperparameters found in the search.

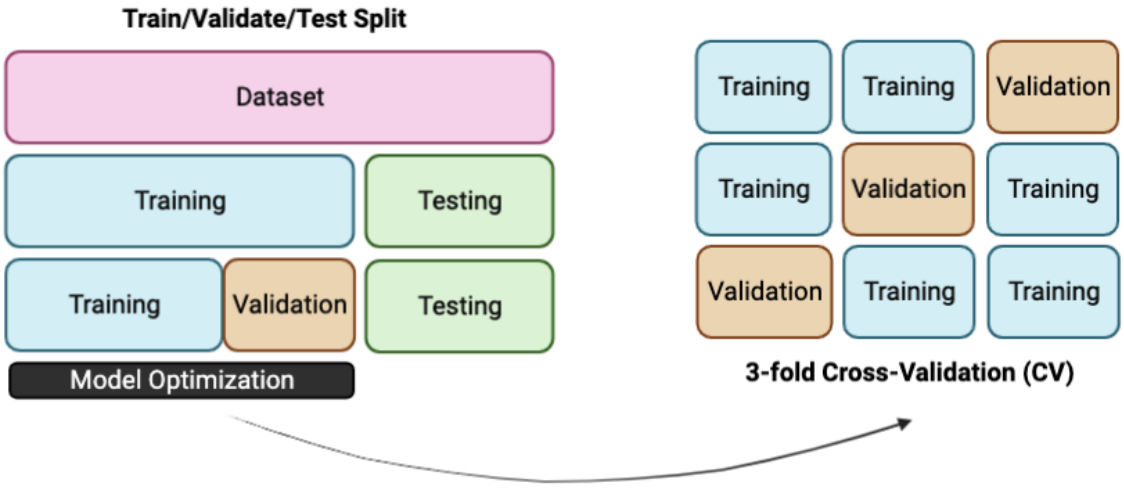

Figure 7. Training/validation/testing split and 3-fold cross-validation (CV)

4 Results

4.1 Time Series Analysis with SMOTE and PCA

Time series analysis of the ECG data using SMOTE to rebalance the dataset and PCA for dimensionality reduction produced very high values for specificity, but very low values for sensitivity and AUROC (Table 9).

Table 9. Time series analysis with SMOTE

Experiment	*Sens.*	*Spec.*	*AUROC*	*F1 Score*	*PPV*	*NPV*
PCA - 90% of variance conserved	1.5	96.5	43.9	2.86	30.0	49.5
PCA - 80% of variance conserved	2.5	95.5	43.5	4.7	35.7	49.4

4.2 Time Series Analysis without SMOTE and PCA

Time series analysis of the ECG data using a manually rebalanced dataset and PCA at 90% and 80% of data variance maintained for dimensionality reduction (Tables 10,11). These experiments produced high sensitivity, low specificity and moderate AUROC score in experiments that included patient metadata. The best performing model on sensitivity uses features with 90% explained variance, while the best performing model on AUROC uses features with 80% explained variance.

Table 10. Time Series analysis without SMOTE, 90% explained variance with PCA

Experiment	*Sens.*	*Spec.*	*AUROC*	*F1 Score*	*PPV*	*NPV*
PCA - 90% of variance conserved + Metadata	84.7	50.6	73.3	72.4	63.2	76.8
PCA - 90% of variance conserved	48.8	63.5	58.0	52.7	57.2	55.3

Table 11. Time series without SMOTE, 80% explained variance with PCA

Experiment	*Sens.*	*Spec.*	*AUROC*	*F1 Score*	*PPV*	*NPV*
PCA - 80% of variance conserved + Metadata	84.1	52.3	74.8	72.6	63.8	76.7
PCA - 80% of variance conserved	7.6	89.4	51.1	12.9	41.9	49.2

4.3 FREQUENCY DOMAIN ANALYSIS

Frequency domain analysis of the ECG data using a manually rebalanced dataset and the Fast Fourier Transform to convert the time series signal into a frequency domain representation, optimized for both AUROC and F1 score (Table 12). Superior models consistently included patient metadata, and the model optimized for AUROC (including patient metadata) yielded the highest sensitivity across experiments.

Table 12. Frequency domain analysis, optimized for AUROC and F1 Score

Experiment, Optimization	*Sens.*	*Spec.*	*AUROC*	*F1*	*PPV*	*NPV*
With Metadata, AUROC	88.2	55.9	76.2	75.9	66.7	82.6
Without Metadata, AUROC	40.6	75.2	63.8	49.1	62.1	55.9
With Metadata, F1	75.8	63.5	76.5	71.4	67.5	72.5
Without Metadata, AUROC	87.1	19.4	64.0	65.1	51.9	60.0

4.4 AUTOMATIC ECG INTERVAL FEATURE ANALYSIS

Analysis of fourteen automatically extracted, interval-based features of the ECG data using a manually rebalanced dataset (Table 13). The best model across all metrics included automatically extracted ECG features and patient metadata, but an increase in sensitivity and specificity and a decrease in AUROC were observed when excluding age and BMI from the patient metadata.

Table 13. Automatic interval based ECG feature analysis

Experiment	*Sens.*	*Spec.*	*AUROC*	*F1*	*PPV*	*NPV*
ECG features	66.5	56.5	67	63.3	60.4	62.8
ECG features + Metadata	73.5	71.8	78.4	72.9	72.2	73.1
ECG features (no axis) + Metadata	45.3	84.7	73.3	56.4	60.8	74.7
Metadata	71.1	70.0	75.6	70.8	70.4	70.8
ECG features + SEX	76.4	57.1	71.4	69.7	64.0	70.8
Sex	71.7	52.9	62.3	65.6	60.4	65.2

The best performing models were those that used the frequency-based features with patient metadata, as well as interval-based features with metadata (Table 14). The frequency-based model yielded a superior sensitivity, but a higher AUROC was observed in the interval-based model.

Table 14. Summary of best performing models

Experiment	*Sens.*	*Spec.*	*AUROC*	*F1*	*PPV*	*NPV*
FFT with Metadata	88.2	55.9	76.2	75.9	66.7	82.6
Automatic Interval-Based ECG features + Metadata	73.5	71.8	78.4	72.9	72.2	73.1

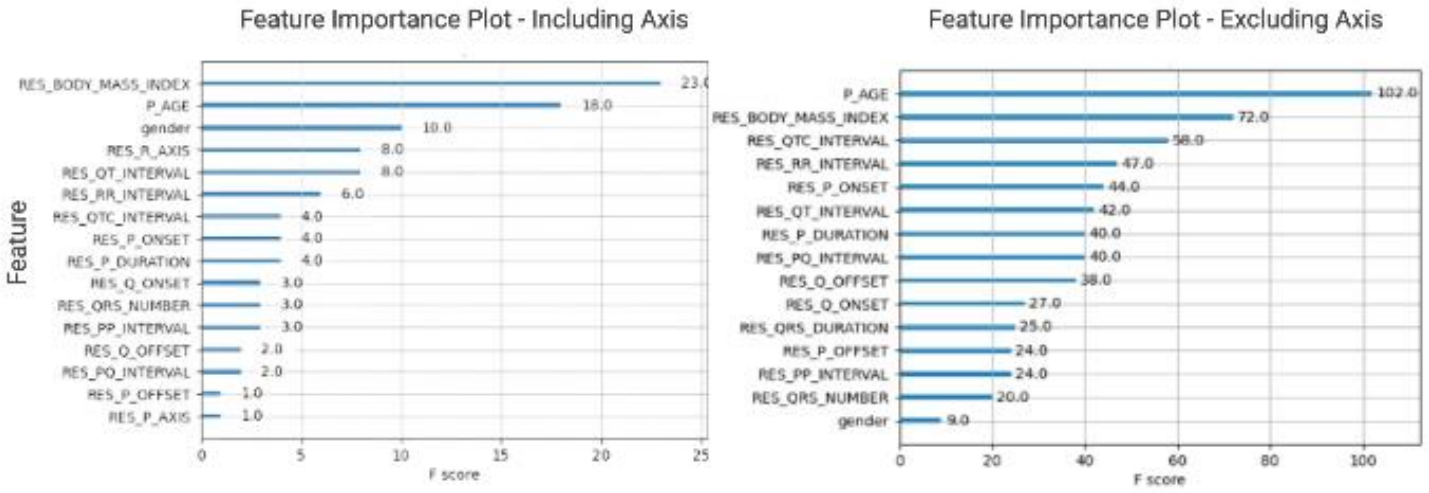

Figure 8. Feature importance plots for automatic interval-based features, with and without axis.

5 Discussion

Overall, the results provide interesting insights into the use of time series data with XGBoost, as well as the utility of different pre-processing and class rebalancing techniques such as Principal Component Analysis (PCA), the Fast Fourier Transform (FFT), and Synthetic Minority Oversampling Technique (SMOTE). Additionally, we leveraged a large dataset (N=7463) of patients with 12-Lead ECG readings to train the models. The two best models used frequency-based and interval-based features as input, achieving an AUROC of **0.784** and sensitivity of **0.882** (Table 14). The frequency-based features were extracted using the FFT on the pre-processed time-series data, while the interval-based features (Table 7) were automatically extracted by the ECG machine. The frequency-based features were larger in number (400 features), while interval-based features were smaller (14 features), but contained clinically explainable features, aiding in the interpretation of the results and demonstrating superior performance on AUROC. The feature-based model, however, performed better on sensitivity. Overall, these findings suggest that it is possible to achieve similar results to state-of-the-art models without the deep learning (DL) approaches they typically employ. The ability to achieve results with XGBoost that are in line with neural network-based models is ideal for several reasons. First, less data is typically required for training XGBoost in comparison to DL models. It is generally accepted that DL models require large volumes of data to train the model accurately due to the very large number of hyperparameters that need

tuning (58). The rule of thumb for the number of training samples required for DL models is that the number of samples should be an order of magnitude higher than the number of features (59). In the dataset, each patient has 5000 ECG features (datapoints), and taking the most conservative form of this rule with an order of magnitude of two, this would require a minimum of 10,000 ECGs. In practice, however, orders of magnitude above two (often 10) are typically used, meaning that minimum number of samples required to produce accurate results expands exponentially. Additionally, training time and inference time may be more efficient with XGBoost compared to DL models, as XGBoost was designed to be highly efficient, leading to decreases in the time required to train a model and the time required to make predictions. This is well-demonstrated by the results of this work, where all the training and predictions were made on a MacBook Pro 2018. In contrast, most DL models require expensive Graphic Processing Units (GPUs) and Central Processing Units (CPUs) for training and making predications, and with large datasets may require multiple units to conduct distributed training. With more efficient models such as XGBoost, infrastructure costs can be reduced, and scalability of the models becomes increasingly accessible.

We found positive results using training and testing datasets that were completely independent with no patient overlap. This is especially important as some previous studies partition the ECG into individual heartbeats from the same patient, treating them as independent samples (60,61). Therefore, a single patient's data may be present in both

the training and testing datasets, however, no single heartbeat can be present in both. In theory, this is a good approach for increasing the number of samples for training the model (a 10 second electrocardiogram could contain 10-16 beats at a normal heart rate). However, in practice, this means a model would be learning representations of the disease on a heartbeat of a patient and being tested on a different heartbeat from the same patient. While this may lead to favourable results, the model may simply be learning representations specific to patients in the training set. To validate the use of these models in real-world scenarios, testing with an independent dataset of different patients is required. Typically, however, all models will see a decrease in performance when being tested on unseen data. This is one of the integral questions in machine learning model development – how well does this model generalize? Generalization refers to the model's ability to make accurate predictions on new data. For this reason, we decided to take each patient's ECG reading in its entirety as a single sample, ensuring that no patients were present in both training and testing datasets. By doing this, we increase the likelihood that the model will generalize appropriately as it is being tested on data of unseen patients. We compare our results to that of the literature, where the ECG-partitioning technique is often used (Table 15).

Table 15. Comparison of results between the literature and the present study.

Authors	*Model Architecture*	*AUROC*	*Sensitivity (%)*	*Specificity (%)*
Lin et. al	Deep Learning	82.55	N/A	N/A
Dave et. al	Random Forests	86.5	79	79
Nguyen et. al	Artificial Neural Network	N/A	70.59	65.38
Present Work	XGBoost: Frequency-Based Features	78.4	73.5	71.8
	XGBoost: Interval-Based Features	76.2	88.2	55.9

Time Series Analysis:

In the time series experiments, the best AUROC and sensitivity achieved were **0.743** and **0.847,** respectively, in the experiments which used PCA as the form of dimensionality reduction (Tables 10 and 11). The best performing model for AUROC was trained on PCA features which conserved 90% of the data original variance (Table 10). The best performing model for sensitivity was trained on PCA features that conserved 80% of the original variance.

The use of SMOTE appeared to negatively impact model performance (Table 9). We hypothesize that the use of this technique introduced noise, and the artificial ECGs generated by it lacked the time-dependent features that relate heart function to diabetes

diagnosis. SMOTE is not a time-series specific method for up-sampling data, and therefore treats each data-point as independent in time. This, in turn, may result in the loss of the time-dependent features important for the model to make predictions. This finding raises the question of how to address limitations around sample size in datasets for models such as those outlined in this work. Training models on larger datasets of real patient data would be ideal but is not always feasible due to constraints around privacy, data-sharing, missing data, and patient consent. During the development of this model, Microsoft released a paper about a novel approach to imbalanced datasets for time series classification problems: Temporal-oriented Synthetic Minority Oversampling Technique for Imbalanced Time Series Classification (T-SMOTE) (62). In this work, they propose an oversampling tool based on the original SMOTE that preserves the temporal information of time series data. This is achieved by the generation of "near-border" samples, meaning samples that are near the border of the two classes being examined – in this case, the presence or absence of T2DM. Intuitively, samples near the class border will contribute more to classification model performance than those which are far from the border. From these near-border samples, additional samples are generated. Of all the near border samples (synthetic and original), a subset is randomly selected to construct the final, class-balanced dataset. In future iterations of this work, T-SMOTE should be applied to examine the effects of time-series specific oversampling techniques, and may provide a

means to augment medical datasets, leading to the development of more robust models that may generalize more effectively.

In the time series experiments that did not employ any oversampling techniques (Tables 10 and 11), we opted to manually rebalance the dataset by under-sampling the majority (T2DM negative) class. The best model in these experiments yielded a AUROC of 0.748 and a sensitivity and specificity of 0.841, and 0.523, respectively when using PCA features that conserved 90% of the data's variance ($N_{features}$ = 236). In experiments using PCA features that conserved only 80% of the data's variance ($N_{features}$ = 162), the sensitivity and AUROC of the model significantly dropped. This suggests that a threshold for PCA variance of 80% may be too low to produce acceptable results. In the context of a screening tool, it is important that the model has adequate sensitivity and specificity. The sensitivity of a model is influenced by false negatives, in this case, patients that the model classifies as disease-negative who in fact, do have the disease. In any screening strategy, it is imperative that we ensure as many diseases positive cases are being detected as possible, and mistakes from the model in this context would mean delayed diagnosis of T2DM. Specificity, on the other hand, is influenced by false positives, or patients that the model classifies as disease positive who in fact, do *not* have the disease. A low specificity in a model may lead to unnecessary additional testing for patients who do not have the disease. This does not pose any significant health risks for patients; however, it would lead to unnecessary healthcare spend. Therefore, we would

favour optimizing sensitivity in models for disease screening, however we must ensure that this doesn't come at the cost of a very low specificity. Additionally, the PPV and NPV of the models describes the likelihood that a patient is diabetes positive given a positive prediction from the model, and vice versa. The best performing model on these metrics used PCA features that conserved 90% of the data's variance, resulting in a PPV and NPV of 0.632 and 0.768, respectively. Overall, in the context of population-level screening, it would be advisable to optimize for and select models that perform best based on sensitivity to ensure no disease-positive patients are missed, in our case, this would be the frequency based model.

Frequency Domain Analysis:

The frequency domain experiments examined the use of frequency-based features as input to the XGBoost model using the Fast Fourier Transform (FFT) to decompose the signal into its frequency components (Table 12). The output of an FFT is the amplitude of each frequency component that comprise the signal and gives information as to which frequencies are most prominent in a time-series. Converting a signal to the time series domain does not result in information loss, rather, it is simply a time-independent representation of the signal. The results of the frequency domain experiments were superior to that of the time-series experiments, with an AUROC, sensitivity and specificity of **0.759, 0.882, and 0.559**, respectively. A model that was more balanced in terms of sensitivity and specificity was also derived, and the resulting AUROC,

sensitivity and specificity were 0.713, 0.758, and 0.635, respectively. Further, the PPV of this model was 0.675. Despite the reduction in sensitivity (from 0.882 to 0.758), we observed an increase in specificity, meaning this model may still be accurate enough as a screening tool without incurring unnecessary costs related to a high number of false positives. Further, if a patient did test positive when screened with the second model, this would correspond to a 67.5% chance that they truly have T2DM. We suspect that models from this analysis outperformed the time-series based models as the technique for dimensionality reduction and feature extraction was more appropriate for use with time-series data. While PCA is a widely used and suitable method for dimensionality reduction, it is inherently a statistical method, and was not designed for use with time-series data. The FFT, however, is designed to describe time-series data in the context of frequencies, making it a potentially more ideal way to reduce dimensionality for time series data for the use of machine learning.

Automatic ECG Features:

We examined the use of automatic, interval-based ECG features as input (Table 7), as they provide a more explainable interpretation of how the model makes predictions when detecting T2DM. Each feature relates to a different element of the cardiac cycle, and the direct link to the functional and structural aspects of the heart are well understood.

The use of the 14 automatically extracted, interval-based features as input to the model produced an AUROC, sensitivity, specificity and PPV of **0.784, 0.735, 0.718, and**

0.722, respectively (Table 13). It is important to note that two of the fourteen features can only be derived from a standard 12-Lead ECG. For this reason, we conducted an additional experiment with the features that can be derived from a single lead (12 features in total) and observed a decrease in sensitivity down to 0.453 and PPV down to 0.608. This suggests these features may be important in making correct predictions. From the feature importance plot of the experiment that included all 14 features, it can be observed that second to patient metadata, R-axis, QT-interval, QTC-interval, and RR-interval were most important when classifying patients (Figure 8). The R-axis is related to ventricular activity, and it has been found that patients with T2DM have a higher prevalence of left ventricular hypertrophy (LVH) (37). LVH typically develops due to hemodynamic overload (related to high blood pressure) and may be a strong predictor of cardiovascular disease in the T2DM population (63). QT-interval and QTC-interval both examine ventricular repolarization (where QTC-interval is the heart-rate adjusted QT-interval), or in other words, the time is takes for the heart to contract and then repolarize in preparation for the next heartbeat. Due to the effect of heart rate on QT-interval, the QTC interval is more commonly used in diagnosing various pathologies. Prolonged QT/QTC-intervals are associated with sudden death and may be observed in some T2DM patients secondary to cardiac autonomic neuropathy (CAN) and diabetic autonomic neuropathy (DAN) (64) (37). Finally, the RR-interval represents the duration of the cardiac cycle and is directly related to heart rate. Elevated heart rate may be observed in T2DM patients

with autonomic dysfunction or due to increased levels of insulin in the blood (hyperinsulinemia) (65).

. We observed similar trends in feature importance when excluding axis-related features, and interestingly, all ECG features were more important than gender in classifying patients (Figure 8). In future work, the exploration of the use of different combinations of ECG features may be informative as to what the minimum number of features are for the detection of T2DM from a single-lead ECG.

A systematic review of the diagnostic accuracy of HbA1c testing (compared to the Oral Glucose Tolerance Test – OGTT) found that the sensitivity and specificity of HbA1c testing ranged from 24%-78% and 79%-100% respectively at a threshold of HbA1c=6.5% for diagnosis (66). The two best performing models had sensitivities of 73.5% and 88.2%, and specificities of 55.9% and 71.8%. Therefore, compared to one of the gold-standards for screening and diagnosis, we have demonstrated results with increased sensitivity. Some models produced higher specificities at 87.4% (using ECG features with no axis and patient metadata), however a decrease in sensitivity down to 45.3% was observed. Nonetheless, both results fall within the ranges observed in the systematic review. Therefore, we have demonstrated comparable results using only Lead-I of a 12-Lead ECG and an easily scalable machine learning model. As discussed in earlier sections the CANRISK tool is typically used to assess risk for T2DM and facilitate additional testing for patients at high risk for the disease. In a study validating the CANRISK model in a

multi-ethnic population, the area under the receiver operating cure (AUROC) was found to be 0.75 (95% CI: 0.73-0.78) (32). The best model in this work yielded an AUROC of 78.4, demonstrating superior performance to the CANRISK questionnaire-based tool. In an additional study that examined the effectiveness of CANRISK in First Nations and Métis in Canada, the sensitivity and specificity of the tool was 68% and 63% among individuals aged 40 or over and 27% and 87%, respectively, among those under 40 (67). In addition to achieving results above those cited in this study for sensitivity and specificity, it should be noted that the CANRISK tool was not developed for those under 40, and the drop in sensitivity in this age group further solidifies the need for flexible, adaptive tools for screening that are effective across larger demographics.

5.1 LIMITATIONS

The study is limited by the dataset in which we have trained, validated, and tested the model. To develop a more robust model that performs well across the general population, the dataset used for training should be as diverse as possible, and reflective of the population of interest. In the case of T2DM and the proposed use-case of an early screening protocol, the model should be further validated in a dataset that considers the sex, gender, and socio-demographic effects of T2DM. Further, we are limited by the parameters that were used for model optimization. Theoretically, a better model may be produced by doing a more thorough optimization of hyperparameters, which will be conducted in subsequent studies.

6 Conclusion

The findings are in line with the current literature, which predominantly use deep-learning approaches. However, the model architecture (XGBoost) is not deep learning based and is more efficient to train and deploy as a result. Further, the use of FFT based features for input was found to be superior to time series-based features of the ECG data. Finally, the best performing model was that which used the interval-based ECG features in combination with patient age, sex, and body mass index. Additionally, we have demonstrated results on par with and surpassing current screening strategies such as the CANRISK model and HbA1c testing. In future studies, the validation of the model with an external dataset is required to verify the results in this work. Further, a prospective study should be conducted to validate the use of the algorithm in a more diverse patient population, with an emphasis on socio-demographic groups that are disproportionately affected by T2DM. Additionally, studies should be conducted to examine the differences between the use of single-lead data from a 12-Lead ECG, and single-lead data from a wearable device, such as a smart watch. Due to the efficient nature of the algorithm, a proposed real-world use case would be for the automated screening of T2DM using single-lead data from a wearable device. In this case, the patient would be able to self-screen for the disease by simply taking an ECG reading on their wearable, and running an application on the device which executes the algorithm. This application would determine whether the patient is at high risk for T2DM and if they should proceed to

confirmational testing. This use case may be increasingly relevant in remote communities and LMICs, where access to standard labs and healthcare overall is limited.

In summary, we have demonstrated a successful application of XGBoost for the detection of T2DM from ECG data and produced results in line with both the current literature around deep learning approaches, as well as the current gold-standard for screenings. The potential for the use of AI-based tools to improve screening protocols for diseases such as T2DM is great and presents an opportunity to revolutionize our current screening strategies.

References

1. IDF_Atlas_10th_Edition_2021.pdf [Internet]. [cited 2023 Aug 3]. Available from: https://diabetesatlas.org/idfawp/resource-files/2021/07/IDF_Atlas_10th_Edition_2021.pdf

2. Diabetes Canada Clinical Practice Guidelines Expert Committee. Diabetes Canada 2018 Clinical Practice Guidelines for the Prevention and Management of Diabetes in Canada. Can J Diabetes; 2018.

3. Bilandzic A, Rosella L. The cost of diabetes in Canada over 10 years: applying attributable health care costs to a diabetes incidence prediction model. Health Promot Chronic Dis Prev Can Res Policy Pract. 2017 Feb;37(2):49–53.

4. Harris MI, Klein R, Welborn TA, Knuiman MW. Onset of NIDDM occurs at least 4-7 yr before clinical diagnosis. Diabetes Care. 1992 Jul;15(7):815–9.

5. Sriram RD, Reddy SSK. Artificial Intelligence and Digital Tools: Future of Diabetes Care. Clin Geriatr Med. 2020 Aug 1;36(3):513–25.

6. McCowen KC, Smith RJ. DIABETES MELLITUS | Classification and Chemical Pathology. In: Caballero B, editor. Encyclopedia of Human Nutrition (Second Edition) [Internet]. Oxford: Elsevier; 2005 [cited 2023 Aug 3]. p. 543–51. Available from: https://www.sciencedirect.com/science/article/pii/B0122266943000788

7. Galicia-Garcia U, Benito-Vicente A, Jebari S, Larrea-Sebal A, Siddiqi H, Uribe KB, et al. Pathophysiology of Type 2 Diabetes Mellitus. Int J Mol Sci. 2020 Aug 30;21(17):6275.

8. Rahman MS, Hossain KS, Das S, Kundu S, Adegoke EO, Rahman MdA, et al. Role of Insulin in Health and Disease: An Update. Int J Mol Sci. 2021 Jun 15;22(12):6403.

9. Wilcox G. Insulin and Insulin Resistance. Clin Biochem Rev. 2005 May;26(2):19–39.

10. Ritzel RA, Michael DJ, Butler PC. Insulin Secretion. In: Henry HL, Norman AW, editors. Encyclopedia of Hormones [Internet]. New York: Academic Press; 2003 [cited 2023 Aug 3]. p. 384–90. Available from: https://www.sciencedirect.com/science/article/pii/B0123411033001789

11. Nakrani MN, Wineland RH, Anjum F. Physiology, Glucose Metabolism. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 [cited 2023 Aug 3]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK560599/

12. Insulin, Glucagon, and Diabetes Mellitus - ClinicalKey [Internet]. [cited 2023 Aug 3]. Available from: https://www-clinicalkey-com.proxy3.library.mcgill.ca/#!/content/book/3-s2.0-B9780323597128000795?scrollTo=%23hl0000232

13. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. J Epidemiol Glob Health. 2020 Mar;10(1):107–11.

14. Reed J, Bain S, Kanamarlapudi V. A Review of Current Trends with Type 2 Diabetes Epidemiology, Aetiology, Pathogenesis, Treatments and Future Perspectives. Diabetes Metab Syndr Obes Targets Ther. 2021 Aug 10;14:3567–602.

15. Chen L, Magliano DJ, Zimmet PZ. The worldwide epidemiology of type 2 diabetes mellitus--present and future perspectives. Nat Rev Endocrinol. 2011 Nov 8;8(4):228–36.

16. Hu FB. Globalization of Diabetes. Diabetes Care. 2011 Jun;34(6):1249–57.

17. Nguyen NT, Nguyen XMT, Lane J, Wang P. Relationship between obesity and diabetes in a US adult population: findings from the National Health and Nutrition Examination Survey, 1999-2006. Obes Surg. 2011 Mar;21(3):351–5.

18. Zhou X, Ji L, Ran X, Su B, Ji Q, Pan C, et al. Prevalence of Obesity and Its Influence on Achievement of Cardiometabolic Therapeutic Goals in Chinese Type 2 Diabetes Patients: An Analysis of the Nationwide, Cross-Sectional 3B Study. PLoS ONE. 2016 Jan 4;11(1):e0144179.

19. Institute for Health Metrics and Evaluation (IHME). GBD Compare [Internet]. University of Washington, Seattle, WA; 2015. Available from: http://vizhub.healthdata.org/gbd-compare

20. Institute for Health Metrics and Evaluation (IHME) [Internet]. Seattle, WA: IHME, University of Washington; 2015. Available from: http://vizhub.healthdata.org/gbd-compare

21. Stress and Type 2 Diabetes: A Review of How Stress Contributes to the Development of Type 2 Diabetes | Annual Review of Public Health [Internet]. [cited 2023 Aug 3]. Available from: https://www-annualreviews-org.proxy3.library.mcgill.ca/doi/10.1146/annurev-publhealth-031914-122921?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed

22. Donath MY, Shoelson SE. Type 2 diabetes as an inflammatory disease. Nat Rev Immunol. 2011 Feb;11(2):98–107.

23. www.heart.org [Internet]. [cited 2023 Aug 13]. Diabetes Risk Factors. Available from: https://www.heart.org/en/health-topics/diabetes/understand-your-risk-for-diabetes

24. DiabetesCanadaWebsite [Internet]. [cited 2023 Aug 13]. Assess your risk of developing diabetes. Available from: https://www.diabetes.ca/type-2-risks/risk-factors---assessments

25. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002 Feb 7;346(6):393–403.

26. Long Term Effects of a Lifestyle Intervention on Weight and Cardiovascular Risk Factors in Individuals with Type 2 Diabetes: Four Year Results of the Look AHEAD Trial. Arch Intern Med. 2010 Sep 27;170(17):1566–75.

27. Gregg EW, Chen H, Wagenknecht LE, Clark JM, Delahanty LM, Bantle J, et al. Association of an intensive lifestyle intervention with remission of type 2 diabetes. JAMA. 2012 Dec 19;308(23):2489–96.

28. Diabetes remission of bariatric surgery and nonsurgical treatments in type 2 diabetes patients who failure to meet the criteria for surgery: a systematic review and meta-analysis | BMC Endocrine Disorders | Full Text [Internet]. [cited 2023 Aug 15]. Available from: https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-023-01283-9

29. Hale I. Add to cart? Can Fam Physician Med Fam Can. 2015 Nov;61(11):937–9, 941–4.

30. Andrade JG, Shah A, Godin R, Lanitis T, Kongnakorn T, Brown L, et al. Cost-effectiveness of atrial fibrillation screening in Canadian community practice. Heart Rhythm O2. 2022 Nov 17;4(2):103–10.

31. Lemieux CL, deGroh M, Gibbons L, Morrison H, Jiang Y. A Tool to Assess Risk of Type 2 Diabetes in Canadian Adults. Can J Diabetes. 2020 Jul;44(5):445–7.

32. Robinson CA, Agarwal G, Nerenberg K. Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population. Chronic Dis Inj Can. 2011 Dec;32(1):19–31.

33. Glycated haemoglobin (HbA1c) for the diagnosis of diabetes. In: Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus: Abbreviated Report of a WHO Consultation [Internet]. World Health Organization; 2011 [cited 2023 Aug 3]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK304271/

34. Kaul P, Chu LM, Dover DC, Yeung RO, Eurich DT, Butalia S. Disparities in adherence to diabetes screening guidelines among males and females in a universal care setting: A population-based study of 1,380,697 adults. Lancet Reg Health Am. 2022 Oct;14:100320.

35. Wilmot E, Idris I. Early onset type 2 diabetes: risk factors, clinical impact and management. Ther Adv Chronic Dis. 2014 Nov;5(6):234–44.

36. Hillier TA, Pedula KL. Complications in young adults with early-onset type 2 diabetes: losing the relative protection of youth. Diabetes Care. 2003 Nov;26(11):2999–3005.

37. The ECG in Diabetes Mellitus | Circulation [Internet]. [cited 2023 May 24]. Available from: https://www-ahajournals-org.proxy3.library.mcgill.ca/doi/full/10.1161/CIRCULATIONAHA.109.897496

38. Lin CS, Lee YT, Fang WH, Lou YS, Kuo FC, Lee CC, et al. Deep Learning Algorithm for Management of Diabetes Mellitus via Electrocardiogram-Based Glycated Hemoglobin (ECG-HbA1c): A Retrospective Cohort Study. J Pers Med. 2021 Jul 27;11(8):725.

39. Dave D., Vyas K., Branan K., McKay S., DeSalvo D.J., Gutierrez-Osuna R., et al. Detection of Hypoglycemia and Hyperglycemia Using Noninvasive Wearable Sensors: ECG and Accelerometry. J Diabetes Sci Technol [Internet]. 2022;((Dave, Erraguntla) Wm Michael Barnes '64 Department of Industrial and Systems Engineering, Texas AM University, College Station, TX, United States). Available from: https://journals.sagepub.com/home/DST

40. Nguyen LL, Su S, Nguyen HT. Neural network approach for non-invasive detection of hyperglycemia using electrocardiographic signals. Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf. 2014;2014:4475–8.

41. Gupta K, Bajaj V. A Robust Framework for Automated Screening of Diabetic Patient Using ECG Signals. IEEE Sens J. 2022 Dec;22(24):24222–9.

42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems [Internet].

Curran Associates, Inc.; 2017 [cited 2023 Aug 3]. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

43. Detection of Hypoglycemia and Hyperglycemia Using Noninvasive Wearable Sensors: ECG and Accelerometry - Darpit Dave, Kathan Vyas, Kimberly Branan, Siripoom McKay, Daniel J. DeSalvo, Ricardo Gutierrez-Osuna, Gerard L. Cote, Madhav Erraguntla, 2022 [Internet]. [cited 2023 Aug 3]. Available from: https://journals-sagepub-com.proxy3.library.mcgill.ca/doi/10.1177/19322968221116393?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed

44. Demystifying Maths of Gradient Boosting | by Krishna Kumar Mahto | Towards Data Science [Internet]. [cited 2023 Aug 7]. Available from: https://towardsdatascience.com/demystifying-maths-of-gradient-boosting-bd5715e82b7c

45. Descending into ML: Training and Loss | Machine Learning | Google for Developers [Internet]. [cited 2023 Aug 7]. Available from: https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss

46. He Z, Lin D, Lau T, Wu M. Gradient Boosting Machine: A Survey [Internet]. arXiv; 2019 [cited 2023 Aug 14]. Available from: http://arxiv.org/abs/1908.06951

47. How XGBoost Works - Amazon SageMaker [Internet]. [cited 2023 Aug 15]. Available from: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

48. Feature dimensionality reduction: a review | Complex & Intelligent Systems [Internet]. [cited 2023 Aug 15]. Available from: https://link-springer-com.proxy3.library.mcgill.ca/article/10.1007/s40747-021-00637-x

49. Roelofs R, Shankar V, Recht B, Fridovich-Keil S, Hardt M, Miller J, et al. A Meta-Analysis of Overfitting in Machine Learning. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2019 [cited 2023 Aug 15]. Available from: https://papers.nips.cc/paper_files/paper/2019/hash/ee39e503b6bedf0c98c388b7e8589aca-Abstract.html

50. Bashir D, Montanez GD, Sehra S, Segura PS, Lauw J. An Information-Theoretic Perspective on Overfitting and Underfitting [Internet]. arXiv; 2020 [cited 2023 Aug 15]. Available from: http://arxiv.org/abs/2010.06076

51. Gallagher NB. Savitzky-Golay Smoothing and Differentiation Filter.

52. Maćkiewicz A, Ratajczak W. Principal components analysis (PCA). Comput Geosci. 1993 Mar 1;19(3):303–42.

53. Almuwaqqat Z, O'Neal WT, Hammadah M, Lima BB, Bremner JD, Soliman EZ, et al. Abnormal P-Wave Axis and Myocardial Ischemia Development During Mental Stress. J Electrocardiol. 2020;60:3–7.

54. Al-Akchar M, Siddique MS. Long QT Syndrome. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 [cited 2023 Aug 13]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK441860/

55. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002 Jun 1;16:321–57.

56. XGBoost Parameters — xgboost 1.7.6 documentation [Internet]. [cited 2023 Aug 13]. Available from: https://xgboost.readthedocs.io/en/stable/parameter.html

57. Refaeilzadeh P, Tang L, Liu H. Cross-Validation. In: LIU L, ÖZSU MT, editors. Encyclopedia of Database Systems [Internet]. Boston, MA: Springer US; 2009 [cited 2023 Aug 15]. p. 532–8. Available from: https://doi.org/10.1007/978-0-387-39940-9_565

58. Deep Learning: A Primer for Radiologists | RadioGraphics [Internet]. [cited 2023 Aug 7]. Available from: https://pubs.rsna.org/doi/abs/10.1148/rg.2017170077?journalCode=radiographics

59. Google for Developers [Internet]. [cited 2023 Aug 7]. The Size and Quality of a Data Set | Machine Learning. Available from: https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality

60. Cordeiro R, Karimian N, Park Y. Hyperglycemia Identification Using ECG in Deep Learning Era. Sensors. 2021 Sep 18;21(18):6263.

61. Kulkarni A.R., Patel A.A., Pipal K.V., Jaiswal S.G., Jaisinghani M.T., Thulkar V., et al. Machine-learning algorithm to non-invasively detect diabetes and pre-diabetes from electrocardiogram. BMJ Innov. 2022;9(1):32–42.

62. Zhao P, Luo C, Qiao B, Wang L, Rajmohan S, Lin Q, et al. T-SMOTE: Temporal-oriented Synthetic Minority Oversampling Technique for Imbalanced Time Series Classification. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence [Internet]. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization; 2022 [cited 2023 Aug 7]. p. 2406–12. Available from: https://www.ijcai.org/proceedings/2022/334

63. Mohan M, Dihoum A, Mordi IR, Choy AM, Rena G, Lang CC. Left Ventricular Hypertrophy in Diabetic Cardiomyopathy: A Target for Intervention. Front Cardiovasc Med [Internet]. 2021 [cited 2023 Aug 7];8. Available from: https://www.frontiersin.org/articles/10.3389/fcvm.2021.746382

64. Vinik AI, Maser RE, Mitchell BD, Freeman R. Diabetic Autonomic Neuropathy. Diabetes Care. 2003 May 1;26(5):1553–79.

65. Hyperinsulinemia produces both sympathetic neural activation and vasodilation in normal humans. - PMC [Internet]. [cited 2023 Aug 7]. Available from: https://www-ncbi-nlm-nih-gov.proxy3.library.mcgill.ca/pmc/articles/PMC296986/

66. Kaur G, Lakshmi PVM, Rastogi A, Bhansali A, Jain S, Teerawattananon Y, et al. Diagnostic accuracy of tests for type 2 diabetes and prediabetes: A systematic review and meta-analysis. PloS One. 2020;15(11):e0242415.

67. Gina A, Ying J, Susan RVK, Chantal L, Heather O, Yang M, et al. Effectiveness of the CANRISK tool in the identification of dysglycemia in First Nations and Métis in Canada. Health Promot Chronic Dis Prev Can Res Policy Pract. 2018 Feb;38(2):55–63.