جمهورية العراق

وزارة التعليم العالي والبحث العلمي
جامعة ديالى
كلية العلوم
قسم علوم الحاسبات

نظام تصنيف دقيق لمرض السكري بالاعتماد على خوارزميات التعليم الالي

رسالة مقدمة
الى كلية العلوم في جامعة ديالى وهي جزء من متطلبات نيل
شهادة الماجستير في علوم الحاسبات

تقدم بها الطالبة

رشا مهدي عبد القادر

بإشراف

الاستاذ المساعد الدكتور عبد الباسط كاظم شكر

1443 هـ **2021 م**

Chapter One

Introduction

Chapter One
Introduction

1.1 Introduction

Diabetes is a long-term condition that occurs when the pancreas is unable to generate enough insulin or when the body does not utilize insulin correctly. Insulin is a hormone that facilitates the regulation of blood sugar levels. Diabetic patients are more likely to have high blood sugar, which causes many different systems in the body to be damaged over time, especially the neurons and blood vessels [1]. Diabetes may cause a variety of severe long-term complications, including heart attack, blood vessels, heart disease, renal failure, peripheral artery disease, stroke, and nerves [2].

According to research by the International Diabetes Federation, the number of individuals living with diabetes is continuously rising. Diabetes sufferers were 382 million in 2013, and 595 million are predicted by 2035 [3].

Diabetes is divided into three categories. The inability of the pancreas to generate enough insulin causes Type 1 Diabetes Mellitus Disease (DMD). The cause has yet to be determined. Second, Type 1 DMD starts with insulin resistance, a condition in which cells fail to react properly to insulin. It's possible that as diabetes progresses, you'll run out of insulin. Excessive body weight and a lack of exercise as result of a changing lifestyle are two frequent reasons. Third, gestational diabetes, which develops during pregnancy in a woman who has never had diabetes before, causes high blood sugar levels to appear abruptly.

Computer science and high-performance computing have aided almost all fields, including medical research, in achieving better outcomes than conventional practical methods [4]. Data mining has already become an essential method to help academics extract knowledge from vast and complicated data, such as patient medical records, in the current digital age [5].

In Diabetes Mellitus research, using machine learning and data mining techniques is a major way to utilize huge quantities of accessible diabetes-related data for extracting knowledge [6]. Machine learning data mining methods are useful in healthcare because they provide a mechanism for analyzing medical data and diagnosing illnesses. For medical professionals and researchers, successfully detecting diabetes is a crucial medical problem [7]. Finding ways to combat the illness, which includes using tools and techniques developed in the area of computer science, is an important subject in medical research [8].

This thesis presents a system to detection and classification Diabetes disease using five most popular and effective machine learning algorithms, as well as comparing the performance of these algorithms used to predict diabetes disease.

1.2 Related Works

Many researchers have proposed many related works about an Accurate Diabetes Classification System based on Machine learning Algorithms. The following are some studies and researchers that can so far associate works to the proposed model of this thesis:

- **M. T. Islam, et al., (2019)** [9] Collected 340 cases with 26 features of patients who previously had diabetes and had a variety of symptoms, which were divided into two categories: Typical and Non-Typical. Cross-validation was used to train the dataset, and three Machine Learning (ML) methods, such as Bagging, LR and RF, were employed for classification. Bagging has an accuracy of 89.12 %, LR has an accuracy of 83.24 %, and RF has an accuracy of 90.29 %.

- **A. Choudhury, and D. Gupta, (2019)** [10] Provide a thorough comparative analysis of different machine learning algorithms for the PIMA Indian Diabetic dataset. All classification methods' accuracy rates, such as LR, decision tree (DT), NB, KNN, and SVM, is evaluated in the performance analysis. It was discovered that logistic regression provides the best accurate results for classifying diabetes and nondiabetic samples, with accuracy rate of 77.6%.
- **O. Daanouni, et al., (2019)** [11] Used four Machine Learning algorithms to predict patients with or without type 2 diabetes mellitus (K-Nearest Neighbors, Decision Tree, Artificial Neural Network, and Deep Neural Network). On two diabetes datasets, these methods were trained and tested: The first obtained from Frankfurt Hospital in Germany (2000 records), while the second came from the Pima Indian (768 records) dataset, which is famous for its quality. All of these databases include a mixture of variables, from risks to clinical information. Data with missing values and noisy data (with or without pre-processing) are both found in the datasets were used to assess the performance of the tested methods (after pre-processing). When the findings are evaluated using various similarity metrics like Accuracy, Sensitivity, and Specificity, the results showed that the results are the best in the KNN algorithm with 97.53%.
- **H. Kaur & V. Kumari (2020)** [12] Created five different models to diagnose diabetes using linear kernel support vector machine (SVM-linear), redial basis kernel, (SVM- Radial Basis Function (RBF)), (KNN), Artificial Neural Network (ANN), and Multifactor Dimensionality Reduction (MDR)

algorithm. The dataset's features are chosen using the Boruta wrapper algorithm to feature selection, which provides unbiased selection of different features. Accuracy, precision, recall, Area Under Curve (AUC), F1 score, and all parameters that were took into account while evaluating the models. In comparison to the other models employed, the experimental findings showed that all of the models produced excellent results; however, The SVM-linear model was showed to have the highest accuracy and precision for diabetes prediction, at 89.0% to (SVM-linear) and 88.0% to (KNN). Compared to the KNN model, the model with the best recall and F1 score had an F1 of 90.0% and an F0 of 88.0%.

- **K. G., Naveen, et al., (2020)** [13] Utilized the criteria Glucose, Blood Pressure, Skin Thickness, Insulin, and Age to identify diabetes. The healthcare industry generates a large number of statistical units. Those data sets are a compilation of diabetic patient information from hospitals. Big data analytics is a kind of processing that analyzes data units and displays hidden information. This information comes from the National Institute of Diabetes and Digestive Diseases' Pima Indians Diabetes Database (PIDD). The dataset's goal is to determine if a patient has diabetes or not, mainly using diagnostic measures from the dataset. The enormous database was used to extract several records. SVM, RF, DT, KNN, and LR, and are the five machine learning algorithms utilized. They found that Random Forest had a higher accuracy rate of almost 75%.
- **N. K. Anwar & R. Saian (2020)** [14] Used two different diabetes datasets the Frankfurt Germany diabetes (2000 records) dataset and the Pima Indian diabetes(768 records) dataset, various machine learning model involved in this study like Naïve Bayes, KNN. The main algorithm that will be used in this research is Ant-Miner to make a comparison in term of accuracy value. The

highest accuracy obtained for the dataset is when will implement Ant-Miner algorithm which is 73.64% compared to other algorithms.

- **L. Miao, et al. (2020)** [15] Seek to create a long-term risk assessment tool for type-2 diabetes T2D related cardiovascular disease (CVD). To construct the prediction models, the Framing-ham Heart Study (FHS) dataset was utilized to train SVM and KNN algorithms. The original dataset was balanced via synthetic Minority Oversampling Technique algorithm then the SVM algorithm was used to train the model, the accuracy of diagnosing CVD in patients with T2D after tuning the parameters and training the model 1000 times was 96.92%, with an 89.8% recall rate. Similarly, the dataset was trained using the KNN method, with a 92.9 % recall rate. Data processing, model training, feature selection, and evaluation are briefly described in this section. They examined the publicly available datasets from four large institutions: FHS, Hospital Frankfurt Diabetes Center, and National Institute of Diabetes and Digestive and Kidney Diseases.

1.3 Problem statement

Diabetes is among critical diseases and lots of people are suffering from this disease. The main issue to detect diabetes based on automated diagnostic and detection systems in its early stage is an accuracy. According to the most recent papers, the accuracy is still up to date.

In addition, it is possible to rely on automated diagnostic and detection systems to detect diabetes in its early stages, but these systems may suffer from defects in design and methods or implementation that negatively affect the accuracy of disease diagnosis in its early stages.

1.4 Aims of thesis

This thesis aims to propose a system that has the ability for early diabetes disease detection based on several machine learning algorithms. This following aim can be achieved by the following objectives:

1- Proposed and design an automatic Diabetes Diagnose system using five classification algorithms of machine learning to successfully detect diabetes in the early stage.

2- Using a random forest algorithm to select the features is most effective in predicting diabetes.

3- Using techniques and algorithms to analyze and process dataset data to obtain the best accuracy in diagnosing diabetes.

1.6 Layout of thesis

The rest of the thesis chapters are clarified as follow:

Chapter two: The theoretical background of the work

Chapter Three: the system's components and architecture

Chapter Four: The results and Evaluation of the Experimental

Chapter Five: The conclusions and future work.

الخلاصة

مرض السكري هو مرض يصيب العديد من الناس في جميع أنحاء العالم. تتزايد معدلات الإصابة به بشكل مثير للقلق كل عام. إذا لم يتم علاجها ، فقد تصبح المضاعفات المرتبطة بالسكري في العديد من أعضاء الجسم الحيوية قاتلة. يعد الاكتشاف المبكر لمرض السكري مهمًا جدًا للعلاج في الوقت المناسب والذي يمكن أن يوقف المرض من التقدم إلى مثل هذه المضاعفات.

لذلك ، حظيت دقة الكشف عن مرض السكري وتشخيصه باهتمام كبير ، كما أن تصميم وتطوير نظام تشخيص يمكنه التعرف على مرض السكري بنجاح في مرحلة مبكرة هو قضية رئيسية للمجتمع العلمي. تحتوي أنظمة التشخيص الحالية على العديد من العيوب، بما في ذلك الحسابات المقعدة واختيار الأدوات أو التقنيات أو الخوارزميات الغير فعالة للحصول على دقة تشخيص عالية. اذن، هناك حاجة إلى نظام دقيق لاكتشاف مرض السكري وتشخيصه باستخدام خوارزميات التعلم الالي الأكثر كفاءة لتجنب مثل هذه العيوب.

في هذا العمل ، تم اقتراح نظام آلي لتشخيص مرض السكري يعتمد على خوارزمية التعلم الآلي. علاوة على ذلك ، يميز هذا العمل نفسه عن السابق من خلال استخدام مرض السكري الكامل. مجموعة بيانات تحتوي على 2000 عينة لـ 9 ميزات. كما أن لديها القدرة على اكتشاف وتشخيص مرض السكري في مرحلة مبكرة. من ناحية أخرى ، تم تنفيذ أكثر تقنيات التعلم الآلي فاعلية للكشف عن مرض السكري وتشخيصه وهي K-Nearest Neighbors ، Logistic Regression ، و Naive Bayes ، و Support Vector Machine ، و Random Forest.

ومع ذلك ، وفقًا للنتائج التي تم الحصول عليها ، لوحظ أن النظام المقترح حقق نتيجة ممتازة بدقة تصل إلى 99٪ مع Random Forest,خلال المقارنة مع K-Nearest Neighbor التي حققت دقة 98.75٪ ، Support Vector Machine الذي حقق دقة 81٪ ، Logistic Regression الذي حقق دقة 77.50٪ ، و Naive Bayes الذي حقق دقة 77.25٪. كما تم مقارنة أداء النظام المقترح مع الأعمال ذات الصلة وحقق أعلى دقة.