Republic of Iraq

Ministry of Higher Education

And Scientific Research

University of Diyala

College of Science

Computer Science Department

# Diagnosis of Diabetes Mellitus Based on New Dataset for Diyala-Baquba City

A Thesis

Submitted to the Computer Science Department \College of Science \University of Diyala
In a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

*By*
*Ahmed Sami Jaddoa*

*Supervised By*

*Prof. Dr. Ziyad Tariq Mustafa Al-Ta'i*

2021 A.D. 1443 A.H.

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

﴿ اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ * خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ * اقْرَأْ
وَرَبُّكَ الْأَكْرَمُ * الَّذِي عَلَّمَ بِالْقَلَمِ * عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ ﴾

صدق الله العلي العظيم

سورة العلق
من الأية (1- 5)

# Acknowledgments

First and foremost, I would like to thank Allah SWT for his blessing and mercy who has guided me in completing this thesis. Then I would like to thank my supervisor, **Prof. Dr. Ziyad Tariq Mustafa Al_Ta'i,** professor of computer science at Diyala University - collage of science, for the great effort he exerted, I would like to thank him for his valuable guidance and support through his supervision of this work.

My thanks to all academics and administrative staff at the Department of computer science.

Last and not least, thanks a lot go to my family, my friends, and anyone who helped me in one way or another.

Ahmed Sami Jaddoa

# Dedication

I would like to dedicate this

Work To:

The owner of a fragrant biography and an enlightened thought: He had the first credit for my obtaining higher education (my beloved father), may God prolong his life.

To the one who set me on the path of life, and made me calm (my dear mother), may God prolong her life.

To my dear brothers and sister who had a great impact on many obstacles and difficulties.

**Ahmed Sami Jaddoa**

# ***Linguistic Certification***

This is to certify that this thesis entitled ***"Diagnosis of Diabetes Mellitus Based on New Dataset for Diyala-Baquba City"*** was prepared by ***"Ahmed Sami Jaddoa"*** at the University of Diyala/ Computer Science Department, is reviewed linguistically. Its language was amended to meet the style of the English language.

**Signature:**

**Name: Dr. Ghazwan Mohammed Jaafar**

**Date: / / 2021**

# Scientific Certification

I certify that the thesis entitled "***Diagnosis of Diabetes Mellitus Based on New Dataset for Diyala-Baquba City***" was prepared by "***Ahmed Sami Jaddoa***" has been evaluated scientifically; therefore, it is suitable for debate by the examining committee.

**Signature:**

**Name:** **Assist. Prof. Dr. Shaima Hamid Shaker**

**Date: / / 2021**

# Scientific Certification

I certify that the thesis entitled "***Diagnosis of Diabetes Mellitus Based on New Dataset for Diyala-Baquba City***" was prepared by "***Ahmed Sami Jaddoa***" has been evaluated scientifically; therefore, it is suitable for debate by the examining committee.

**Signature:**

**Name: Prof. Dr. Belal Ismail Khalil Ibrahim**

**Date: / / 2021**

# Supervisor’s Certification

We certify that this thesis entitled “***Diagnosis of Diabetes Mellitus Based on New Dataset for Diyala-Baquba City”*** was prepared by **"*Ahmed Sami Jaddoa*"** Under our supervisions at the University of Diyala, Faculty of Science, Computer Science Department, as partial fulfillment of the requirement needed to award the degree of Master of Science in Computer Science.

**(Supervisor)**

**Signature:**

**Name: Prof. Dr. Ziyad Tariq Mustafa Al_Ta'i**

**Date: / / 2021**

Approved by the University of Diyala Faculty of Science Department of Computer Science.

**Signature:**

**Name: Assist. Prof. Dr. Bashar Talib Al-Nuaimi**

**Date: / / 2021**

**(Head of Computer Science Department)**

# Examination Committee Certification

We certify that we have read the thesis entitled "***Diagnosis of Diabetes Mellitus Based on New Dataset for Diyala-Baquba City***" and an examination committee examined the student "***Ahmed Sami Jaddoa***" in the thesis content and that in our opinion, it is adequate as fulfill the requirement for the Degree of Master of Science in Computer, University of Diyala.

**(Chairman)**
**Signature:**
**Name: Prof. Dr. Taha Mohammed Hassan**
**Date: / / 2021**

**Signature:**
**Name: Prof. Dr. Abbas Fadhil Mohammed Ali** **(Member)**
**Date: / / 2021**

**Signature:**
**Name: Assist. Prof. Dr. Abdulbasit Kadhim Shukur** **(Member)**
**Date: / / 2021**

**Signature:**
**Name: Prof. Dr. Ziyad Tariq Mustafa Al_Ta'i** **(Supervisor)**
**Date: / / 2021**

Approved by the **Dean** of College of Science, University of Diyala

**(The Dean)**
**Signature:**
**Name: Prof. Dr. Tahseen H. Mubarak**

**Date: / / 2021**

# Abstract

Diagnosing diabetes and pre-diabetes early has a great level of importance, to provide the patients with the ability for managing the disease early and possibly delay or prevent the serious complications of the disease, which may result in decreasing the quality of life. It may be helpful in the reduction of the risks of serious disease developments, like premature heart diseases and stroke, limb amputation, blindness, and renal failure.

In the Proposed System, a system is proposed to diagnosis diabetes mellitus. The proposed system is based on the Chi-square test, Information gain, and a new hybrid method for feature selection. The new hybrid method is proposed to reduce the number of features to a minimum number by intersecting the Chi-square test and Information gain methods. The results of feature selection are fed into the classification stage to obtain the best accuracy. Five classification algorithms are utilized: Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression (LR) to classify absence or presence of diabetes mellitus disease.

The proposed system is tested using Precision, Specificity, Sensitivity, f-score, and Accuracy. The results of the proposed system have experimented on two datasets (Local and Global (Pima)). Algorithms (RF, NB, SVM, KNN, and LR) achieved maximum accuracy (98%) with a hybrid method, while these algorithms achieved accuracy between (94% and 98%) with Chi-square test and Information gain on Local dataset. Algorithms (LR and NB) achieved maximum accuracy (91.17%) with a hybrid method, while (KNN) achieved accuracy (85.29%) and (RF, SVM) achieved accuracy (86.76%). Algorithms (RF, NB, KNN, LR, and SVM) achieved accuracy between (79.41% and 89.70%) with Chi-square test and Information gain on Global (Pima) dataset.

# Lists of Contents

| **Contents** | | ***Page No*** |
|---|---|---|
| Abstract | | I |
| Lists of Contents | | II |
| List of Figures | | IV |
| List of Tables | | V |
| List of Algorithms | | VI |
| List of Abbreviations | | VII |
| | ***Chapter One: Introduction*** | **1-11** |
| 1.1 | Introduction | 1 |
| 1.2 | Related Work | 3 |
| 1.3 | Problem Statement | 10 |
| 1.4 | Aims of the Thesis | 10 |
| 1.5 | Contribution | 11 |
| 1.6 | Outline of Thesis | 11 |
| | ***Chapter Two: Theoretical Background*** | **12-36** |
| 2.1 | Introduction | 12 |
| 2.2 | Data Mining | 12 |
| 2.3 | Data Preprocessing | 13 |
| 2.3.1 | Data Cleaning | 13 |
| 2.3.1.1 | Missing Values | 14 |
| 2.3.1.2 | Data Transformation | 14 |
| 2.4 | Feature Selection | 16 |
| 2.4.1 | Chi-square Test | 17 |
| 2.4.2 | Information Gain | 18 |
| 2.5 | Classification Algorithms | 19 |
| 2.5.1 | Random Forest Algorithm (RF) | 20 |
| 2.5.2 | Naïve Bayes Algorithm (NB) | 22 |
| 2.5.3 | Support Vector Machine Algorithm (SVM) | 25 |
| 2.5.4 | K-Nearest Neighbor Algorithm (KNN) | 29 |
| 2.5.5 | Logistic Regression Algorithm (LR) | 31 |
| 2.6 | Evaluation Measurements of Classification | 35 |
| | ***Chapter Three:*** **The Proposed System Design** | **37-49** |
| 3.1 | Introduction | 37 |
| 3.2 | Diabetes Mellitus Disease Dataset | 37 |
| 3.3 | The Proposed Diagnosis System | 39 |

| | | |
|---|---|---|
| 3.3.1 | Data Preprocessing | 41 |
| 3.3.2 | Feature Selection Stage | 42 |
| 3.3.2.1 | Chi-square Test Feature Selection | 42 |
| 3.3.2.2 | Information Gain Feature Selection | 43 |
| 3.3.2.3 | Hybrid Feature Selection | 44 |
| 3.3.3 | Classification | 45 |
| 3.3.3.1 | Random Forest Algorithm | 45 |
| 3.3.3.2 | Naïve Bayes Algorithm | 46 |
| 3.3.3.3 | Support Vector Machine Algorithm | 47 |
| 3.3.3.4 | K-Nearest Neighbor Algorithm | 48 |
| 3.3.3.5 | Logistic Regression Algorithm | 48 |
| | ***Chapter Four: Results and Discussion*** | **50-78** |
| 4.1 | Introduction | 50 |
| 4.2 | Tools and Database | 50 |
| 4.3 | Results of the Proposed System | 50 |
| 4.3.1 | Results of Data Preprocessing Stage | 50 |
| 4.3.2 | Results of Feature Selection Stage | 52 |
| 4.3.3 | Results of Classification Stage | 55 |
| 4.3.4 | Graphical Representation of Results | 69 |
| 4.4 | Comparison to Related Works | 75 |
| 4.5 | Discussion | 76 |
| | ***Chapter Five: Conclusions and Suggestions For Future Works*** | **79-80** |
| 5.1 | Introduction | 79 |
| 5.2 | Conclusions | 79 |
| 5.3 | Suggestions for Future Work | 80 |
| | ***References*** | **81-87** |

# List of Figures

| *Caption* | *Page No.* |
|---|---|
| **Figure (2.1):** Data Mining is the Main Stage in the KDD process | 13 |
| **Figure (2.2):** The Process of FS | 17 |
| **Figure (2.3)**: The Process of Classification | 20 |
| **Figure (2.4):** Random Forest Simplified | 22 |
| **Figure (2.5):** SVM training with two classes | 28 |
| **Figure (2.6):** Fundamental design of KNN classifier | 30 |
| **Figure (2.7):** The Logistic Function | 32 |
| **Figure (2.8):** The two-class LR model is a general linear model with a logistic link function P | 33 |
| **Figure (3.1):** Block Diagram of the Proposed System Design | 40 |
| **Figure (4.1):** Result Chi-square Test using Local dataset | 69 |
| **Figure (4.2):** Results of Information gain using Local Dataset | 69 |
| **Figure (4.3):** Result of Chi-square Test using Pima Indian Diabetes dataset | 70 |
| **Figure (4.4**): Results of Information gain using Pima Indian Diabetes Dataset | 70 |
| **Figure (4.5):** Results of Accuracy of the Methods of Classification without the use of the Feature Selection on Local Dataset | 71 |
| **Figure (4.6):** Results of the classification methods accuracy using Chi-Square Test Feature Selection on Local Dataset. | 71 |
| **Figure (4.7):** Accuracy of classification methods using Local Dataset with the Feature Selection of the Information Gain | 72 |
| **Figure (4.8):** Accuracy results of the approaches of classification using Local Dataset with the Hybrid Feature Selection | 72 |
| **Figure (4.9):** Classification Methods Accuracy Results utilizing Pima Indian Diabetes dataset without Feature Selection | 73 |
| **Figure (4.10):** Accuracy Results of Classification Approaches using Pima Dataset with Chi-square Test Feature Selection | 73 |
| **Figure (4.11):** Accuracy Results of Classification Approaches using Pima Dataset with Information gain Feature Selection | 74 |
| **Figure (4.12):** Accuracy results of Classification Approaches using Pima Dataset with Hybrid Feature Selection | 74 |

# List of Tables

| *Caption* | *Page No* |
|---|---|
| **Table (1.1):** The Summary of the Related Work | 8 |
| **Table (2.1):** Confusion Matrix | 36 |
| **Table (3.1):** Local Dataset Attribute Description | 37 |
| **Table (3.2):** Attribute Description of PIDD | 38 |
| **Table (4.1):** Summary of Local dataset | 51 |
| **Table (4.2):** Summary of Pima Indian Diabetes dataset | 52 |
| **Table (4.3):** Chi-square Test Results of Local Dataset | 53 |
| **Table (4.4):** Information gain Results of Local Dataset | 54 |
| **Table (4.5):** Chi-square Test Results of Pima Indian Diabetes Dataset | 54 |
| **Table (4.6):** Information gain Results of Pima Indian Diabetes Dataset | 55 |
| **Table (4.7):** Results of classifier algorithms for Local dataset without Feature Selection | 56 |
| **Table (4.8):** Results of classifier algorithms for Local dataset with Chi-square Test Feature Selection | 57 |
| **Table (4.9):** Results of classifier algorithms for Local dataset with Information gain Feature Selection | 58 |
| **Table (4.10):** Results of classification algorithms for Local dataset with Hybrid Feature Selection | 60 |
| **Table (4.11):** Results of classification algorithms for Pima Indian Diabetes dataset without Feature Selection | 62 |
| **Table (4.12):** Results of classification algorithms for Pima Indian Diabetes dataset with Chi-square Test Feature Selection | 63 |
| **Table (4.13):** Results of classification algorithms for Pima Indian Diabetes dataset with Information gain Feature Selection | 65 |
| **Table (4.14):** Results of classification algorithms for Pima dataset with Hybrid Feature Selection | 67 |
| **Table (4.15):** Comparison of the Suggested Work with Other Works | 75 |

# List of Algorithms

| ***Caption*** | ***Page No.*** |
|---|---|
| **Algorithm (2.1):** Random Forest Algorithm | 21 |
| **Algorithm (2.2):** Steps Model Classifier of Naive Bayes Algorithm | 23 |
| **Algorithm (2.3):** Naïve Bayes Classification  Algorithm | 24 |
| **Algorithm (2.4):** SVM-based Classifier | 28 |
| **Algorithm (2.5):** K-Nearest Neighbor Algorithm | 30 |
| **Algorithm (2.6):** Logistic Regression classifier | 33 |
| **Algorithm (3.1):** Data Cleaning Algorithm | 41 |
| **Algorithm (3.2):** Finding Best Feature using Chi-square Test | 43 |
| **Algorithm (3.3):** Finding Best Feature using Information gain | 44 |
| **Algorithm (3.4):** Finding Best Feature using Hybrid Feature | 45 |
| **Algorithm (3.5):** Implemented RF Algorithm | 46 |
| **Algorithm (3.6):** Implemented NB Algorithm | 46 |
| **Algorithm (3.7):** Implemented SVM Algorithm | 47 |
| **Algorithm (3.8):** Implemented KNN Algorithm | 48 |
| **Algorithm (3.9):** Implemented LR Algorithm | 49 |

# List of Abbreviations

| *Abbreviations* | *Meaning* |
|---|---|
| **AB** | Adaptive Boosting |
| **ANN** | Artificial Neural Network |
| BMI | Body Mass Index |
| **C4.5** | C4.5 Decision Trees |
| **C5.0** | C5.0 Decision Trees |
| **DM** | Diabetes mellitus |
| **EM** | Expectation- Maximization Algorithm |
| **FS** | Feature Selection |
| **GA** | Genetic Algorithm |
| **GBT** | Gradient Boosted Trees |
| **IDF** | International Diabetes Federation |
| **IG** | Information Gain |
| **IR** | Iterative Relief |
| **J48** | J48 Decision Trees |
| **JRIP** | JRIP Decision Trees |
| **KDD** | Knowledge Discovery Process |
| **KNN** | K Nearest Neighbors |
| **LR** | Logistic Regression |
| **MLP** | Multi-Layer Perceptron |
| **NB** | Naïve Bayes |
| **NN** | Neural Network |
| **NIDDK** | National Institute of Diabetes and Digestive and Kidney |
| **PCA** | Principal Component Analysis |
| **PIDD** | Pima Indian Diabetes Dataset |
| **RBF** | Radial Basis Function Network |
| **RepTree** | Reduced Error Pruning Tree |
| **RF** | Random Forest |
| **RFE** | Recursive Feature Elimination |
| **SMO** | Sequential Minimal Optimization |
| **SS** | Stability Selection |
| **SVM** | Support Vector Machine |

# Chapter One

# Introduction

## 1.1 Introduction

Nowadays, people face various diseases due to environmental condition and their living habits. Thus, there is high importance in diagnosing diseases at an earlier stage. Nevertheless, it is difficult for doctors to make a precise diagnosis based on only symptoms. For this problem to be solved, data mining is used effectively for diagnosing diseases [1].

Globally, one of the most lethal diseases is Diabetes Mellitus since millions of individuals are affected by it. It is caused by a high-sugar diet and other unhealthy eating and lifestyle choices, like the absence of consistent physical activities. Also, the disease's onset might be caused by genetics [2].

The International Diabetes Federation (IDF) states that in 2019, the worldwide prevalence of diabetes is expected to be 463 million individuals (9.3%), it might increase by 2030 to 578 million (10.2%), while by the year 2045, it might be 700 million (10.9%). In urban, DM is high (10.8%) compared to rural (7.2%) areas; also, its prevalence is less in low-income nations (4.0%) compared to high-income ones (10.4%). Half of the individuals experiencing diabetes do not know that they have diabetes[3].

In addition, diabetes is one of the chronic diseases which is characterized by high levels of blood glucose in the human body. As time goes by, diabetes results in damage to the heart, eyes, kidneys, and so on. Commonly, it is difficult for medical professionals to detect diabetes' early prediction [4].

Diabetes mellitus happens when the human body cells become resistant to insulin or when not enough insulin is produced via the pancreas. The energy that exists in food can't be used effectively by humans due to diabetes [5].

The major diabetes types are **Type-1 diabetes –** in which insulin isn't produced by the body. Early-onset diabetes, juvenile diabetes, and insulin-dependent diabetes are a few other names for this type. Commonly, it occurs in young people and children. Type-1 diabetes constitutes about 10% of all diabetes cases. [6]. **Type-2 Diabetes –** the body cells don't react to insulin (insulin resistance), or insufficient insulin is produced via the body for appropriate function. Commonly, it occurs at the age of 40 years old. This type constitutes about 90% of all worldwide diabetes' cases [6]. **Type-3 Diabetes - Gestational diabetes**: During pregnancy, females are affected by this type. In their blood, a few females have high glucose levels, and not enough insulin is produced by their bodies for all glucose to be transported to cells, leading to progressively increasing glucose levels. In this type, the diagnosis is made throughout pregnancy [6].

Medical data mining can be defined as one of the approaches to find significant patterns which assist in the medical diagnosis, while the process of knowledge extraction is referred to as data mining. □ne of the data mining tasks is data classification [□].

All data set instance is classified via the process of classification into various groups according to the information indicated via its features. Determining the effective features is complicated with no prior knowledge. Thus, many features are typically provided to the data set, which involves redundant, irrelevant, and relevant features. □et, redundant and irrelevant features aren't important for classification□they might even decrease the performance of classification because of the large search space[□].

□ariable selection or feature selection (□□) is utilized for enhancing the data mining algorithms efficiency□□□ methods are utilized with the data. Also, it is a process used to identify and remove maximum redundant and irrelevant information. □ot all available attributes are useful in the database. Commonly, many attributes are obtained, yet just a few of them are utilized. In a real-world problem, many redundant, irrelevant, and noisy features are in the data[9].

In this thesis, a diagnosis system of diabetes mellitus using data mining techni□ues with optimal cost and better performance is proposed.

## 1.2 Related Work

- **K. Thangadurai and N. Nandhini (2016)** [10]: □uggested a system to predict and diagnose diabetes mellitus persons. The system used classification techni□ues, including □A, □□M, C4.5, EM, and □- means to classify diabetes data. The efficiency of the developed model is based on □ima India Diabetes dataset. The suggested approach for records classification with the EM algorithm achieved an accuracy of □0%, whereas, the C4.5 algorithm achieved □1.□% accuracy, □ means achieved □□% accuracy, □□M achieved 6□% accuracy, and the □A achieved □□.1% accuracy.

❖ **K. Saravanapriya and J. Bagyamani (2017)** [11]: Analyzed the performance of the classification techni□ues in the diabetes data set. This model used classification techni□ues such as □□, □4□, □□, M□□, □□I□, □□□, □□M, and □□□ □etwork to be classified for diabetes data. The efficiency of the developed model is based on □ima India Diabetes dataset. The suggested approach for records' classification with □□ achieved an accuracy of □□%, whereas the □4□ algorithm achieved □6% accuracy, □□ algorithm achieved □□% accuracy, □□I□ algorithm achieved □6.5% accuracy, M□□ achieved □4% accuracy, □□□ algorithm achieved □5% accuracy, □upport □ector Machine algorithm achieved □9% accuracy, and the □□□ □etwork algorithm achieved □0% accuracy.

❖ **J. Steffi, D. R. Balasubramanian, and M. K. Aravind Kumar (2018)** [1□]: Introduced a system for □redicting Diabetes Mellitus using Data Mining Techni□ues, which used □□, □□, A□□, C5.0 Decision Tree, and □□M. The efficiency of the developed model is based on □ima India Dataset. The suggested approach for records' classification with □□ achieved an accuracy of □□.5□%, whereas the □□ algorithm achieved □4.6□% accuracy, with (C5.0) achieved an accuracy of □4.6□%, with the (A□□) algorithm achieved □□.□9% accuracy and the (□□M) algorithm achieved □□.1□% accuracy.

- **S. S. Mirzajani and S. Salimi (2018)** [1□]: □roposed a system for diagnosis of DM using data mining which utilizes □□, C5.0, □ayesian network, and □□M, which have been compared to predict diabetes. The efficiency of the developed model was based on □ima India Diabetes dataset. The suggested approach for records classification with C5.0 achieved an accuracy of □0.□%, whereas the □□ algorithm achieved □□.6% accuracy, the □ayesian network algorithm achieved □□.0□% accuracy, and the □□M achieve □□.□□% accuracy.
- **K. Akyol and B. Şen (2018)** [14]**:** Introduced a system to distinguish normal persons or diabetic ones with □major phases. In the first one, the □□ or weighting approaches were examined for finding the most important attributes for the disease where used I□, □□E, and □□ algorithms. In the second step, the performances regarding □□T, A□, and □□ ensemble learning algorithms were assessed. The efficiency of the developed model was based on □ima India Diabetes dataset. □ased on the experimental results, the accuracy of prediction regarding a combination of A□ and □□ approach is somewhat better compared to other algorithms with a classification accuracy of □□.□□%**.**
- **K. M. Varma and Dr. B. S. Panda (2019)** [15]: Compared the performance analysis of □□, □□□, C5.0, and □□M to predict diabetes using Machine □earning Techni□ues. The efficiency of the developed model was based on □ima India Diabetes dataset. The suggested techni□ue for records classification with □□□ achieved an accuracy of □□.5□%, whereas the □□ algorithm

achieved □4.6□% accuracy, the C5.0 algorithm achieved □4.6□% accuracy and the □□M achieved □□.1□% accuracy.

- **M. Warke et al. (2019)** [16]: Introduced a system for Diabetes Diagnosis using Machine □earning Algorithms, which used Decision trees, □□□, □□, and □□M. In addition, the efficiency regarding the development model was based on □ima India Diabetes dataset. The suggested approach for records' classification with □□ achieved an accuracy of □□%, whereas the Decision Tree algorithm achieved 6□% accuracy, the □□M algorithm achieved 6□% accuracy, and the □□□ algorithm achieved 66% accuracy.
- **M. F. Faruque, Asaduzzaman, and I. H. Sarker (2019)** [1□]: □uggested a system for □erformance Analysis of Machine □earning Techni□ues to □redict Diabetes Mellitus, which used □□M, □□, □□□, and C4.5 algorithms. This model used the □ima India Diabetes dataset. The suggested techni□ue for records classification with □□ achieved an accuracy of 6□%, whereas with the C4.5algorithm achieved □□% accuracy, the □□M algorithm achieved □0% accuracy, and the □□□ algorithm achieved □1% accuracy.
- **T. M. Alam et al.(2019)** [1□]: Introduced a system to select considerable attributes through □rincipal Component Analysis (□CA). The results specified that there is a strong relation between glucose levels, □MI, and diabetes, which has been extracted through the Apriori approach. A□□, □□, and □-means clustering approaches have been used to predict diabetes. The efficiency of

the developed model was based □ima India Diabetes dataset. The best accuracy (□5.□%) was recorded by A□□.

- **S. A. Mahmoudinejad Dezfuli et al. (2019)** [19]: Developed an ensemble system with the use of data mining techni□ues based on 4 classification approaches, simple decision tree, □□□, Ensemble method, and □□ algorithms for detecting diabetes mellitus. The efficiency of the developed model was based on □ima India Diabetes dataset. □uch classifiers give □□.0% accuracy for the decision tree, given the accuracy of □□.□0% for a □□□, give the accuracy of □9.□0% for □□ and give the accuracy of □0.60% for the Ensemble method.
- **P. Sonar and K. Jaya Malini (2019)** [□0]: Introduced a system for Diabetes □redication using Different Machine □earning Approaches, which used DT, □□, □□M, and A□□ algorithms. In addition, the efficiency regarding the development system was based on □ima India Diabetes dataset. The suggested approach for records classification with DT algorithm achieved an accuracy of □4%, whereas □□ algorithm achieved □0% accuracy, □□M and A□□ algorithms achieved □□% accuracy.
- **N. Razali et al. (2020)** [□1]: □roposed a system using many techni□ues of data mining like □□, □M□, □epTree, and □imple □□ to classify whether a negative or positive result of diabetes diagnostics. The efficiency of the developed model was based on □ima India Diabetes dataset. These techni□ues gave the accuracy of □□.60% for □□, whereas gave the accuracy of □5.□0% for □imple □□, gave the accuracy of □5.10% for □epTree, and gave the accuracy of □4% for □e□uential Minimal □ptimization (□M□).

- ❖ **L. J. Muhammad, E. A. Algehyne, and S. S. Usman (2020)** [□□]: Introduced a system to □redictive □upervised Machine □earning Models for Diabetes Mellitus, which used □□, □□, □□□, □□M, and □□T algorithms. In addition, the diagnostic dataset for the DM type □ patients was collected from the Murtala Mohammed □pecialist □ospital, □ano □tate, in □igeria. The dataset has nine attributes, including age, family history, glucose, cholesterol (C□□□), blood pressure (□□), □D□ (high density lipoprotein), triglyceride, □MI (body mass index), and the diagnosis result. The dataset has □□□ instances. The suggested approach for records classification with □□ algorithm achieved an accuracy of □□.94%, whereas □□ algorithm achieved □0.□□% accuracy, □□□ algorithm achieved □□.□5% accuracy, □□M algorithm achieved □5.□9% accuracy, and the □□T algorithm achieved □6.□6% accuracy.

Table (1.1) illustrates the summary of the related work.

**Table (1.1):** The □ummary of the □elated □ orks

| **Authors** | **Title** | **Algorithms** | **Accuracy** |
|---|---|---|---|
| □. Thangadurai and □. □andhini (□016) [10] | Comparison of data dining algorithms for predication and diagnosis of diabetes mellitus | □A, □□M, C4.5, EM, and □-means | □□.1% □A high accuracy |
| □. □aravanapriya and □. □agyamani (□01□) [11] | □erformance Analysis of Classification Algorithms on Diabetes Dataset | □□, □4□, □□, M□□, □□I□, □□□, □□M, and □□□ □etwork | □0%□□□ □etwork high accuracy |
| □. □teff, D. □. □alasubramanian, and M. □. Aravind □umar (□01□) [1□] | □rediction of Diabetes Mellitus using Data Mining Techni□ues | □□, □□, A□□, C5.0, and □□M | □4.6□% □□ high accuracy |

| | | | |
|---|---|---|---|
| □. □. Mirzajani and □. □alimi (□01□) [1□] | □redication and Diagnosis of Diabetes by □sing Data Mining Techni□ues | □□, C5.0, □ayesian □etwork, and □□M | □0.□% C5.0 high accuracy |
| □. Akyol and □. □en (□01□) [14] | Diabetes Mellitus Data Classification by Cascading of □eature □election Methods and Ensemble □earning Algorithms | I□, □□E, and □□ □eature □election. A□, □□T, and □□ Classification | □□.□□% A□ with □□ high accuracy |
| □. M. □arma and Dr. □. □. □anda (□019) [15] | Comparative analysis of □redicting Diabetes □sing Machine □earning Techni□ues | □□, □□□, C5.0, and □□M | □4.6□% □□ high accuracy |
| M. □ arke et al. (□019) [16] | Diabetes Diagnosis using Machine □earning Algorithms | Decision Tree, □□□, □□, and □□M | □□%□□ high accuracy |
| M. □. □aru□ue, Asaduzzaman, and I. □. □arker (□019) [1□] | □erformance Analysis of Machine □earning Techni□ues to □redict Diabetes Mellitus | □□M, □□, □□, and C4.5 | □□% C4.5 high accuracy |
| T. M. Alam et al. (□019) [1□] | A model for early prediction of diabetes | □CA □eature □election. A□□, □□, and □-means | □5.□% A□□ best accuracy |
| □.A.Mahmoudinejad Dezfuli et al. (□019) [19] | Early Diagnosis of Diabetes Mellitus □sing Data Mining and Classification Techni□ues | □imple Decision Tree, □□□, Ensemble method, and □□ | □0.60% Ensemble method high accuracy |
| □. □onar and □. □ayaMalini (□019) [□0] | Diabetes prediction using different machine learning approaches | DT, □□, □□M, and A□□ | □□% A□□ method high accuracy |
| □. □azali et al. (□0□0) [□1] | Analyzing Diabetic Data using Classification | □□, □M□, □epTree, and □□ | □5.□0% □epTree high accuracy |
| □. □. Muhammad, E. A. Algehyne, and □. □. □sman (□0□0) [□□] | □redictive □upervised Machine □earning Models for Diabetes Mellitus | □□, □□, □□□, □□M, and □□T | □6.□6% □□T high accuracy |

## 1.3 Problem Statement

Clinical decisions are usually decided depending on the doctor's intuition and expertise instead of the knowledge-rich data hidden in the database. This practice leads to undesired results and high medical costs. The busy style of living people with all the fast food and get back to sit and work, along with less activity and a lack of exercise, has pushed over the edge. These factors boosted the rate of diabetes mellitus disease to a high percentage. Diagnosis of diabetes mellitus disease is a highly risky task because it is affecting directly human life. Accuracy is a factor of high importance because it can be disastrous if not diagnosis accuracy. The diagnosis and incidence of diabetes in Ira□-Diyala were not previously covered. Therefore, diabetes mellitus disease diagnosis is the problem of this thesis.

## 1.4 Aims of the Thesis

The aims:

1- □uilding a diabetes diagnosis system using two types of data sets (□ocal and □lobal) to obtain the best accuracy.

□- □eature selection using (Chi-s□uare test and Information gain). Then using five algorithms for classification: (□andom □orest algorithm (□□), □a□ve □ayes algorithm (□□), □ector Machine support algorithm (□□M), □-□earest □eighbor algorithm (□□□), and □ogistic □egression algorithm (□□)). Then evolution performance of the diabetes diagnosis system.

## 1.5 Contribution

In this study, the major objective is to building a dataset for diabetics in Diyala □overnorate, Ira□. This dataset has been obtained from consulting laboratories at the □a□ubah □eneral □ospital. The second contribution is to building a hybrid method to reduce the number of features in the dataset to a minimum to obtain the important and main features in the diagnosis by comparing the results of the two methods used in selecting the important features and then entering the results into the classification to obtain the best accuracy.

## 1.6 Outline of Thesis

In this study, the other chapters are provided in the following way:

### Chapter Two: Theoretical Background

This chapter gives the background and review of diagnosis diabetes mellitus, feature selection techni□ues, and classification techni□ues.

### Chapter Three: The Proposed System Design

The suggested Diabetes Mellitus diagnosis with its implementation and design is presented in this chapter.

### Chapter Four: Results and Discussion

The evaluation and results obtained from the suggested diagnosis are presented in this chapter.

### Chapter Five: Conclusions and Suggestions for Future work

Conclusions and future work are provided in this chapter.