

DESIGN AND DEVELOPMENT OF A MODEL FOR DIAGNOSIS OF DIABETES MELLITUS

Thesis Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

in

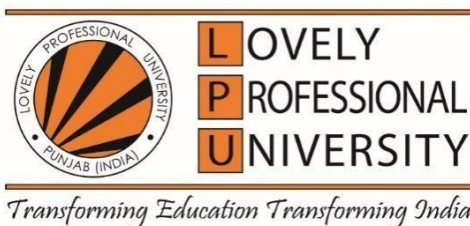
COMPUTER APPLICATIONS

By

OVASS SHAFI ZARGAR

Registration Number: 41800776

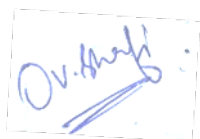
SupervisedBy	Co-Supervisedby
Dr. AVINASH BHAGAT (11002) Department of Computer Applications (Professor) Lovely Professional University, Jalandhar, Delhi G.T Road (NH-1), Phagwara, Punjab (India) – 144411	Dr.TAWSEEF AHMAD TELI Department of Computer Applications (Assistant Professor) Government Degree College for Boys, Anantnag, Jammu and Kashmir



LOVELY PROFESSIONAL UNIVERSITY, PUNJAB
2025

Declaration

I, Ovass Shafi Zargar, hereby declare that the presented work in the thesis entitled “DESIGN AND DEVELOPMENT OF A MODEL FOR DIAGNOSIS OF DIABETES MELLITUS” in fulfilment of degree of Doctor of Philosophy (Ph.D.) is outcome of research work carried out by me under the supervision of Dr. Avinash Bhagat working as Associate Professor, in the Department of Computer Applications of Lovely Professional University, Punjab, India. In keeping with general practice of reporting scientific observations, due acknowledgements have been made whenever work described here has been based on findings of other investigator. This work has not been submitted in part or full to any other University or Institute for the award of any degree.



(Signature of Scholar)

Name of the scholar: Ovass Shafi Zargar

Registration No.:41800776

Department/school: Computer Applications

Lovely Professional University, Punjab, India

Certificate

This is to certify that the work reported in the PhD thesis entitled “DESIGN AND DEVELOPMENT OF A MODEL FOR DIAGNOSIS OF DIABETES MELLITUS” submitted in fulfilment of the requirement for the award of the degree of Doctor of Philosophy (PhD) in the Department of Computer Applications, is a research work carried out by Ovass Shafi Zargar, Registration No.) 41800776, is a bonafide record of his/her original work carried out under my supervision and that no part of the thesis has been submitted for any other degree, diploma or equivalent course.



(Signature of Supervisor)
Supervisor: Dr. Avinash Bhagat
Designation: Associate Professor
Department: Computer Applications
Institution: Lovely Professional University,
Punjab, India



(Signature of Co-Supervisor)
Co-supervisor: Dr. Tawseef Ahmed Teli
Designation: Assistant Professor
Department/: Computer Applications
Institution: Higher Education Department,
J&K, India

Abstract

The work aims to address the critical need for an accurate and timely diagnosis of diabetes mellitus. The research focuses on developing a robust model that can effectively predict diabetes based on relevant medical information. The study utilizes a comprehensive approach, encompassing preprocessing and the implementation of suitable ML techniques, including deep learning techniques.

The study explores the advantages and limitations of existing diagnostic methods and proposes innovative solutions for enhancing the accuracy and efficacy of diabetes prediction. The goal of the study is to advance medical technology and provide valuable viewpoints about the development of effective and accessible tools for early diabetes detection, leading to better disease management and improved patient outcomes.

Diabetes mellitus, a worldwide health emergency, requires early identification to effectively manage the condition and prevent complications. The study investigates how AI is used in the prediction of DM. The research discusses the difficulties in obtaining an early diagnosis, such as vague symptoms, slow development, and restricted access to care. To tackle these issues, the research investigates several data mining methods and how well they predict diabetes, with particular attention on ANN, KNN, and SVM.

The research also focuses on how crucial feature selection and missing value management are to the correctness of the model throughout the data preparation stage. The findings demonstrate the improved capabilities of deep learning models, especially when processing intricate data patterns. The study provides a DNN-based model that uses missing value imputation and feature significance analysis approaches to improve model robustness. The research evaluates the suggested model using the PIMA Indians Diabetes Dataset and contrasts it with conventional machine learning techniques. The results demonstrate that the DNN model performs better than alternative techniques. The findings show how deep learning may improve public health outcomes and have potential for early diabetes prediction.

Deep learning is a highly useful technique for diabetes mellitus early diagnosis. However, using a numerical medical dataset, such as the PIMA Indians Diabetes Database, presents challenges for conventional convolutional neural network (CNN) models. Overcoming this barrier requires a method for converting numerical data into visual representations based on feature significance. Using robust CNN models for early diabetes diagnosis is made possible by this conversion.

The work showcases a new method that facilitates the use of intricate and deep structures for numerical data processing. The PIMA dataset, which has a small amount of data records and is unsuitable for deep learning model training, was used for the majority of the research. The PIMA dataset does not allow for the use of data augmentation and comprises binary data in numerical form. Convolutional Neural Network (CNN)

models' capacity to adjust to numerical inputs makes this possible. Moreover, the incorporation of data augmentation methods is simplified when working with photos linked to diabetes. This makes it possible to build a more reliable training dataset, which might result in DM diagnostic models that are more precise and broadly applicable.

Improvements may be made to diabetes prediction algorithms' accuracy and effectiveness by transforming numerical patient data (blood glucose levels, BMI, etc.) into image data and using Deep Networks which is a core objective of this study. Researchers may take advantage of sophisticated image processing and deep learning techniques by integrating image datasets into diabetes prediction models to determine possible biomarkers and by examining image characteristics, find novel visual biomarkers connected to diabetes and achieve results in more complete and precise diabetes prediction models, which will enhance patient outcomes and healthcare administration.

Acknowledgements

Bowing my head before Almighty Allah, the creator and sustainer of the worlds, I thank Him for his blessings.

The existence of this moment owes a great many thanks to a number of people who helped and provided continuous encouragement and guidance to me, especially my mentor and well-wishers. I hereby take opportunity to express my heartfelt gratitude towards all such people who have been directly or indirectly involved in the success of this endeavor.

The love, encouragement, care and help of my laudable parents are too great to be expressed. I would like to thank my parents who have always been supportive to me and provided me with the resources beyond their capacity.

With profound gratitude, I would like to thank my esteemed supervisor, Dr Avinash Bhagat, Associate Professor, Department of Computer Applications, Lovely Professional University, Phagwara, Punjab for his constant supervision, encouragement and endearment, without which this journey was unattainable.

I would like to extend my deepest gratitude to my co-supervisor, Dr. Tawseef Ahmed Teli, for his invaluable support, guidance, and encouragement throughout the course of this research. Sincere thanks to him for fostering my professional growth and development; his guidance has had a profound impact on my academic journey.

Further, I would like to extend my gleeful thankfulness to the whole faculty of the Department of Computer Applications, Lovely Professional University, Phagwara, Punjab for their humble support and help whenever needed. I would like to express my heartfelt appreciation to Dr. Sophiya Sheikh, Associate Professor, Department of Computer Applications, Lovely Professional University, for her dedication and personal attention to my progress that inspired and motivated me throughout this journey. I also extend my appreciation and thankfulness to all my friends for their support and suggestions at every step of this journey.

Lastly, I would like to thank the non-teaching staff and other members of the department for their cooperation.

Contents

Declaration	1
Certificate	2
Abstract	3
Acknowledgements	5
List of Figures	9
List of Tables	11
Abbreviations	13
1 Outline of the Study	1
1.1 Introduction	1
1.2 Problem Statement	2
1.2.1 Research Significance and Motivation	2
1.3 Research Objectives	3
1.4 Research Gap	4
1.5 Research Contribution	4
1.6 Thesis Organization	5
2 Background and Literature Survey	7
2.1 Introduction	7
2.2 Importance of Early Diagnosis	8
2.2.1 Current Diagnostic Methods	9
2.2.2 Challenges and Limitations	10
2.3 Advancements in ML and DL	10
2.4 Previous Research and Models	11
2.5 Literature Review	12
2.5.1 Methodology for Literature Review	12

2.6	Complexities with Design and Development	18
2.6.1	Challenges in Diabetes Diagnosis	18
2.6.2	Diagnosis using Artificial Intelligence	18
2.6.3	Deep Learning and Numerical Dataset	19
2.7	Comparative Analysis	20
2.7.0.1	Machine and Deep Learning	21
2.7.1	Experiment and Results	44
2.8	Conclusion	45
3	Enhancing Diabetes Mellitus Diagnosis - A Comparative Analysis of Pre-processing Techniques for Data Optimization	48
3.1	Introduction	48
3.2	Conventional machine learning techniques	49
3.2.1	Supervised Learning	49
3.2.1.1	Decision Tree (DT)	49
3.2.1.2	Support Vector Machine (SVM)	50
3.2.1.3	Artificial Neural Network (ANN)	53
3.2.1.4	K-nearest neighbor (KNN)	54
3.2.2	Unsupervised Learning	55
3.2.2.1	Clustering Techniques	55
3.2.2.2	Association Rule Learning	55
3.3	Deep Learning Techniques	55
3.3.1	Convolutional Neural Network	56
3.3.2	CNN Architecture	58
3.3.2.1	The Convolutional Layer	58
3.3.2.2	Pooling Layer	60
3.3.2.3	Fully Connected Layer	61
3.3.2.4	Non-Linearity Layers	61
3.3.2.5	Models of Convolutional Neural Network	62
3.3.2.6	Residual Network (ResNet)	63
3.3.2.7	ResNet Architecture	65
3.3.3	Deep Belief Network	65
3.3.4	Recurrent Neural Network	66
3.4	Experiment and Results without preprocessing	68
3.5	Dataset	68
3.5.1	Advantages	69
3.5.2	Disadvantages	70
3.5.3	Age of Dataset	70
3.6	Preprocessing	71
3.6.1	Data Cleansing	72
3.6.1.1	Managing Missing Values	72
3.6.1.2	Handling Outliers	74
3.6.2	Data Transformation	74
3.6.2.1	Normalization	74
3.6.2.2	Encoding Categorical Variables	75

3.6.3	Feature Selection	76
3.6.3.1	Feature Selection Models	76
3.6.4	Feature Engineering	78
3.6.4.1	Developing Derived Features	78
3.6.4.2	Dimensionality Reduction	79
3.6.5	Normalization of Data	80
3.6.5.1	Handling Skewed Data	80
3.6.6	Data division	80
3.6.6.1	Train-Test Split	80
3.6.7	Managing Imbalanced Data	81
3.7	Experiment and Results with Preprocessing	81
3.8	Classification using DNN	82
3.8.1	Methodology for proposed DNN Model	83
3.8.2	Experiment and Results	84
3.9	Conclusion	86
4	Novel Approach to convert text-based PIMA dataset into image dataset	88
4.1	Introduction	88
4.2	Proposed Methodology	89
4.2.1	Handling of Missing Values	89
4.2.2	Standardization	91
4.2.3	PIMA to Image Dataset	91
4.2.3.1	Original PIMA Dataset	92
4.2.3.2	PIMA Image Dataset	92
4.2.3.3	DNN-based Model using the Image Dataset	94
4.3	Conclusion	98
5	Conclusion and Future Directions	99
5.1	Conclusion	99
5.2	Future Directions	102
A	An Appendix	105
	Bibliography	108

List of Figures

2.1	Methodology for Literature Review	13
2.2	ML Based Publications, ML techniques for Diabetes Diagnosis	22
2.3	DL Based Publications, DL techniques for Diabetes Diagnosis	22
2.4	CM of various ML techniques	45
3.1	Decision Tree classification of PIMA dataset	50
3.2	Linearly Separable Data Points	51
3.3	Multiple Hyper planes	51
3.4	Selecting Hyper plane for data with outlier	52
3.5	Hyperplane which is the most optimized one	52
3.6	Original 1D dataset for Classification	53
3.7	Mapping 1D Data to 2D	53
3.8	KNN algorithm working visualization	54
3.9	The Hierarchical Cortical Model of Cat's Visual Cortex	56
3.10	Representation of image as a grid of pixels	57
3.11	Architecture of CNN	58
3.12	Illustration of Convolution Operation	59
3.13	Convolution Operation	60
3.14	Pooling Operation	61
3.15	Working of LeNet	62
3.16	Network Structure of AlexNet Model	63
3.17	Network Structure of GoogleNet Model	63
3.18	Comparison of 26 layer VS 56 layer architecture	64
3.19	Skip Connection	65
3.21	Deep Belief Network Structure	66
3.22	Recurrent Neural Network	67
3.23	Deep Learning Classification	67
3.24	Description of the PIMA Dataset	69
3.25	Attribute distribution of PIMA Dataset using Histogram	71
3.26	Attribute distribution of PIMA Dataset using Density plots	71
3.27	Feature Selection	76
3.28	Feature Selection Methods	77
3.29	Filter Feature Selection	77
3.30	Wrapper Feature Selection	78
3.31	Intrinsic Feature Selection	78

3.32	Proposed Methodology for a DNN based Model	83
3.33	Proposed Methodology	84
3.34	Feature Selection Performance (%)	85
3.35	Missing Value Performance (%)	85
3.20	ResNet 34 Architecture	87
4.1	Propose Methodology	89
4.2	Methodology for Conversion of Text Pima dataset into Image dataset . .	94
4.3	Resnet50 Model	95
4.4	VGG16 Model	95
4.5	Confusion MatrixResnet50 Model	96
4.6	Confusion Matrix VGG16 Model	96
4.7	Performance Chart VGG16 and Resnet50	97

List of Tables

2.1	Comparative Analysis ML-based Techniques	22
2.1	Comparative Analysis ML-based Techniques	23
2.1	Comparative Analysis ML-based Techniques	24
2.1	Comparative Analysis ML-based Techniques	25
2.1	Comparative Analysis ML-based Techniques	26
2.1	Comparative Analysis ML-based Techniques	27
2.1	Comparative Analysis ML-based Techniques	28
2.1	Comparative Analysis ML-based Techniques	29
2.1	Comparative Analysis ML-based Techniques	30
2.1	Comparative Analysis ML-based Techniques	31
2.1	Comparative Analysis ML-based Techniques	32
2.2	Comparative Analysis DL-based Techniques	33
2.3	Limitations of various ML and DL based Methods	34
2.3	Limitations of various ML and DL based Methods	35
2.3	Limitations of various ML and DL based Methods	36
2.3	Limitations of various ML and DL based Methods	37
2.3	Limitations of various ML and DL based Methods	38
2.3	Limitations of various ML and DL based Methods	39
2.3	Limitations of various ML and DL based Methods	40
2.3	Limitations of various ML and DL based Methods	41
2.3	Limitations of various ML and DL based Methods	42
2.3	Limitations of various ML and DL based Methods	43
2.4	Performance of various ML algorithms	46
2.5	Rates of various ML algorithms	46
3.1	Attributes	68
3.2	Various feature selection techniques	79
3.3	Accuracy (Different Classifiers) without pre-processing Techniques	81
3.4	Artificial Neural Network performance after pre-processing Techniques . .	82
3.5	Application of missing value imputation and normalization on Artificial Neural Network	82
3.6	Performance (%) with and without Feature Selection	84
3.7	Performance (%) with Mean, Median and Polynomial Regression	85
4.1	Image Dataset Division	94
4.2	Performance Comparison ReNet50 and VGG16	97

4.3	Comparative analysis with other DNN-based works	97
A.1	Definitions	105

Abbreviations

ACC	ACCuracy
ADA	AdaBoosted Decision Trees
AdaBoost	Adaptive Boosting
AdR	AdaBoostRegressor
AE	AutoEncoder
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the (ROC) Curve
BDT	Boosted Decision Tree
BILSTM	Bidirectional Long Short-Term Memory
BN	Bayesian Network
BNN	Bayesian Neural Network
BP	BackPropagation
BRNN	Bidirectional Recurrent Neural Network
CART	Classification And Regression Tree
CC	Cloud Computing
CDBN	Convolutional Deep Belief Networks
CLSTM	Convolutional Long Short-Term Memory
CM	Confusion Matrix
CNN	Convolutional Neural Network
ConvNet	Convolutional Neural Network
CTCPN	Convolution Trained Compositional Pattern Neural
DBM	Deep Boltzmann Machine

DL	Deep Learning
DM	Diabetes Mellitus
DNN	Deep Neural Network
DT	Decision Tree
F1 Score	Harmonic Precision-Recall Mean
FC	Fully-Connected
FC-CNN	Fully Convolutional Convolutional Neural Network
FC-LSTM	Fully Connected Long Short-Term Memory
FCN	Fully Convolutional Network
FN	False Negative
FNN	Feedforward Neural Network
FNR	False Negative Rate
FONN	Firefly Optimized Neural Network
FP	False Positive
FPR	False Positive Rate
GD	Gradient Descent
k-NN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LightGBM	Light Gradient-Boosting Machine
LSTM	Long Short-Term Memory
LVQOAC	Learning Vector Quantization Optimized with Ant Colony
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MODLNN	Memetic Optimized Deep Learning Neural Network
MSE	Mean Squared Error
NB	Naive Bayes
NLP	Natural Language Processing
NN	Neural Network
PReLU	Parametric Rectified Linear Unit-Yor Topic Modelling
RBFN	Radial Basis Function Network
RBM	Restricted Boltzmann Machine

ReLU	Rectified Linear Unit
REPTree	Reduced Error Pruning Tree
RF	Random Forest
RL	Reinforcement Learning
RMSE	Root MSE
RNN	Recurrent Neural Network
RNNLM	Recurrent Neural Network Language Model (RNNLM)
ROC	Received Operating Characteristic
Sen	Sensitivity
SGD	Stochastic Gradient Descent
Spec	Specificity
SVM	Support Vector Machine
SVR	Support Vector Regression
TLSTM	Traditional Long Short-Term Memory
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
XGBoost	eXtreme Gradient Boosting

Dedicated to Irhaa and Ehaab...

Chapter 1

Outline of the Study

1.1 Introduction

Among the prevalent, fatal illnesses that have afflicted people worldwide from centuries is Diabetes mellitus, that is indicated by a rise in blood sugar. For human body to work properly, glucose tends to be the most important form of energy but in order to enter into the human cell it needs a hormone produced by pancreas known as insulin. When an individual has the condition diabetes, either the pancreas is not able to release the necessary quantity of insulin or the cells stop responding to the amount of insulin generated. Another form of diabetes is called gestational diabetes, which exclusively affects women who are pregnant and typically goes away on its own after delivery.. Numerous symptoms, including increased appetite and thirst, frequent urination, blurred vision, exhaustion, delayed wound healing, and recurring infections, are linked to diabetes. Diabetes can have a number of major side effects, including heart disease, if it is not properly and promptly treated. The best way to treat diabetes involves changes in the lifestyle and a healthy diet chart, engaging in daily physical activities, exercises along with the medication necessary to control the diabetes like insulin injections, hypoglycemic agents and other medicines required to control the blood sugar levels. Keeping the diabetes under control is one of the important tasks and can be done by continuous monitoring and adjustment of treatment so that the blood sugar levels will be within a healthy range.

1.2 Problem Statement

Among the many chronic illnesses, Diabetes Mellitus pertains to the conditions that is very critical, characterized by higher blood glucose levels. The diabetes can have serious health issues if left untreated. In medical world, in spite of immense technological advancements, timely and accurate diagnosis of diabetes is still a huge challenge. There are number of traditional techniques used to diagnose diabetes but these techniques are either intrusive in nature or lack precision leading to a considerable delay in timely treatment. It is need of an hour to develop an effective and trustworthy technique for early diagnosis of diabetes. A model that might aid with diabetes early detection can utilize developments in medical science, data analytics and artificial intelligence to facilitate early diagnosis of the disease so that customized management approach can be followed for timely control of the disease. The model should consider patient related parameters like demographic data, clinical history, biomarkers and other lifestyle factors.

1.2.1 Research Significance and Motivation

A research study indicated that by 2030 Diabetes Mellitus is projected to affect 537 million persons worldwide irrespective of their age and gender [1]. which is believed to increase to 643,000000. For medical science researchers around the world, Diabetes mellitus is one of the substantial health issues [2]. To effectively manage the disease, a prompt and accurate diagnosis of diabetes is important as the timely action can reduce complications and improve the recovery options of patients [3]. The present techniques used for early diagnosis of diabetes lack accuracy, usability, scalability and reliability underscoring the pressing need for novel alternatives [4]. Creating a model for the early diagnosis of this disease is extremely important in several ways:

1. Enhanced Patient Care: The capability of early identification of persons who are at risk of diabetes mellitus may completely transform patient care [5]. The threat of diseases like cardiovascular problems, neuropathy and retinopathy can be decreased significantly by early detection of diabetes as it makes timely interventions easier to implement like pharmaceutical interventions, lifestyle adjustments and patient education [6].
2. Healthcare Efficiency: The methodology proposed can help to improve healthcare by expediting the diagnostic process and enabling risk classification. The patients with high risk can be given more priority for further assessment and treatment.

Medical practitioners can optimize the delivery of healthcare facilities and reduce costs related to undiagnosed or poorly controlled diabetes [7].

3. **Personalized Medicine:** The Model's capability to incorporate variety of patient related variables can make customized risk assessment and management a reality [8]. The doctors can adapt treatment program to each patients personal needs by considering their lifestyle factors, clinical history and personal variances in genetic predisposition, thereby, optimizing therapeutic efficacy and better treatment adherence.
4. **Research Advancement:** With the development of model for early diagnosis of diabetes, the medical practitioners can understand the disease scientifically and computationally and its direct clinical implications [9]. By utilizing cutting-edge techniques in bioinformatics, data analytics and ML, this research advances the comprehension of the intricate interactions of genetic, environmental and lifestyle factors in the pathogenesis of disease[10].
5. **Impact on Public Health:** Diabetes Mellitus has resulted in tremendous pressure on healthcare system due to its significant socioeconomic cost [1]. The proposed paradigm will provide the opportunity to minimize this burden by helping early diagnosis and efficient management of the disease which could ultimately lower healthcare expenses and will boost productivity and improve overall population health [11].

The importance of addressing the drawbacks of current diagnostic techniques and utilizing strategies that are data-driven to counteract the diabetes pandemic is the primary goal pertaining to this work. With the design of novel and robust model for early diagnosis of diabetes mellitus, the research aims to significantly impact the public health conditions, research and healthcare practice related to DM.

1.3 Research Objectives

The thesis looks at the function of AI, ML, and DL techniques and how they might be used to diagnose DM early. The different issues that arise during the development of model for early diagnosis of diabetes and how those issues can be handled using Deep Learning are the focus areas of the study. The goals are defined as follows:

1. To review various existing models for the diagnosis of diabetes.

2. To perform preprocessing, feature extraction, and feature selection.
3. To prepare a model for diabetes diagnosis using data mining and deep learning.
4. To test and validate the proposed model with the existing models/model.

1.4 Research Gap

1. Many researchers use the PIMA dataset without any pre-processing technique like normalization. As a result, the results suffer from outliers, overfitting, underfitting and other anomalies [12].
2. There have been studies that used limited machine learning algorithms to diagnose diabetes and do not handle the missing values [13].
3. In other studies, some authors have inhibited the application of feature extraction fully [14].
4. Some studies do not consider the importance of all the attributes of the dataset [15]. Features such as height and BMI etc and how they affect the prediction of DM have not been considered, which affects the performance of the classifier [16].
5. Almost all research on Diabetes Prediction has been done on PIMA dataset in textual form which makes it unsuitable for Deep Learning as there are only 768 instances that are not sufficient to train deep neural network models [17].

1.5 Research Contribution

1. Algorithm Comparison: Using the raw PIMA dataset, the study methodically develops and assesses several conventional machine learning (ML) models. This enables comparison with previous studies and offers a standard by which to measure the efficacy of various ML algorithms for DM diagnosis.
2. Effect of Pre-processing and Feature Engineering: The study examines how several pre-processing methods (normalization, standardization, MVI, feature selection, and feature importance) affect these ML models' performance. This emphasizes how crucial data preparation is to improving diagnostic results.

3. **New Image-Based Method Using Deep Learning:** Converting the conventional textual PIMA dataset into an image dataset is a significant achievement in terms of applying DNN-based models for better results. To create a bigger dataset, this innovative method subsequently applies data augmentation techniques, utilizing the capabilities of deep learning (DL).
4. **Promising Outcomes with the Image-Based DL Model:** In comparison to the findings of other researchers and previous studies, the study examines a trained model on this dataset and the possibility of transforming tabular medical data into an image format in order to enhance the performance of deep learning models.
5. **Stressing the Benefits of Deep Learning:** The study specifically highlights the benefits of deep learning over conventional machine learning models, especially when it comes to managing huge datasets and revealing intricate hidden patterns. This supports the investigation of DL methods for diagnosing DM.

The study finds that Random Forests, Gradient Boost, and Logistic Regression work well with conventional ML and recommends that these methods be taken into account for further research. It also highlights the importance of a bigger image-dataset for deep learning-based methods. The work essentially makes a contribution by assessing conventional machine learning techniques as well as developing a new image-based deep learning methodology for DM diagnosis, showcasing its potential for increased accuracy.

1.6 Thesis Organization

Chapter two discusses in detail the background associated with the Diabetes Mellitus and the role of Deep Learning in controlling the chronic diseases like Diabetes Mellitus. Chapter two provides an extensive literature review of uses of the AI, ML and DL techniques in controlling DM and application of various pre-processing techniques for building more reliable and accurate models for timely diagnosis of diabetes mellitus. Chapter three provides a comparative analysis of application of various pre-processing algorithms such as feature selection, imputation of missing values, and feature importance and shows how the performance of model enhances by using pre-processing techniques. It also focuses on the complexities that are associated with design and creation of a model for prediction of DM and the techniques applicable to handle those complexities. The analysis of these methods is also given in this chapter. Chapter four gives a novel approach for converting text based PIMA dataset with small number of records into

Image dataset and application of methods for augmenting data to increase the size of a dataset so that the model can be trained with more data for better prediction and classification. It provides the application of DLL based model on PIMA Image dataset to achieve promising results. Finally, Chapter five provides concluding remarks and future work.

Chapter 2

Background and Literature Survey

2.1 Introduction

Diabetes Mellitus, which is characterized by a rise in blood sugar levels, is the frequent fatal condition that has afflicted population all over the world [18]. Glucose is a vital energy source for the human body to function correctly but in order to enter into the human cell it needs a hormone produced by pancreas known as insulin. When someone has diabetes, their body's cells either stop responding to insulin generated by the pancreas (Type 1 diabetes) or stop responding to insulin altogether (Type 2 diabetes). Gestational diabetes is another form of the disease that only affects women who are pregnant and typically goes away on its own after birth [19]. Diabetes can cause a variety of symptoms, including increased appetite and thirst, frequent urination, blurred vision, exhaustion, delayed wound healing, and recurring infections. Along with drugs like insulin injections, oral hypoglycemic agents, or other treatments to control blood sugar levels, major measures in the treatment of diabetes include making lifestyle changes including eating a nutritious diet and exercising every day [20]. To keep blood sugar levels within a safe range, diabetes management necessitates continuous monitoring and therapy adjustments as needed [21]. Three major categories can be used to describe diabetes:

1. Type 1 Diabetes: This disease is caused due to disorder in autoimmune system of human body where autoimmune system damages the insulin secreting cells and

results in a rise in blood sugar. Insulin pumps and injections are used to treat the condition [22]. Just around 10% of all cases of diabetes pertain to this kind [23].

2. Type 2 Diabetes: This diabetes forms more than 90% of total diabetes cases and is mostly caused by poor life style of people like obesity, poor diet, low exercise and inactive behavior. This type of diabetes is treated with oral medication and in some chronic cases, insulin injections are used in this type of diabetes [24].
3. Gestational Diabetes: Mostly found only in females during the onset of pregnancy period [25]. The body sometimes develops insulin resistance due to hormonal changes during pregnancy that leads to gestational diabetes. People suffering from gestational diabetes may later develop Type 2 diabetes [26].

Some of the common symptoms associated with diabetes mellitus are [24]:

- a. Recurrent Urination
- b. Enhanced Thirst
- c. Unexpected Weight Reduction
- d. Abnormal Hunger
- e. Weariness
- f. Blurry Vision
- g. Slow Healing

2.2 Importance of Early Diagnosis

Diabetes if left untreated for long duration can result in serious problems, including cardiovascular ailments, kidney breakdown, neuropathy, retinopathy, and amputation. The well-known proverb “A stitch in time saves nine” is applicable in early detection of all diseases. Timely cure of diabetes is of utmost importance due to various reasons like:

1. Early Intervention and Management: Early detection of diabetes may help in well-timed intervention and control of the disease for example changing the life style by proper diet control, exercise and proper medical treatment so that the onset of disease can be completely eliminated or delayed [27].
2. Quality of Life: Timely management of diabetes can help in improving the quality of common man’s life by maintaining blood sugar level and avoiding various

complications like kidney failure, cardiovascular problems, hypertension and many more [28].

3. **Cost Savings:** If the diabetes is left undiagnosed and untreated for along time, it may result in multiple diseases that may not only be difficult to control but also much costly [17]. The timely management of diabetes can eliminate such costs and help individuals in leading healthy life without costly medicines and medical procedures [29].
4. **Education and Empowerment:** By early diagnosis of diabetes, patients can get more knowledge about the disease, its management and necessary lifestyle changes required for controlling the disease [30].
5. **Reduced Mortality Risk:** When the disease will be detected at early stage before the onset of complications associated with the disease, proper precautions shall be taken to keep the disease under control which ultimately may result in improvement of healthcare conditions of people and automatically reduces the mortality rate [31].
6. **Research and Development:** The early detection of disease opens new opportunities for researchers to analyse the progression of the disease, possibilities of its prediction before the onset of the disease using artificial intelligence and better cure of the disease so that there will be improvement in the quality of life of common people [32].

2.2.1 Current Diagnostic Methods

In order to diagnose the diabetes mellitus some of the traditional used are:

1. **Fasting test:** With this test, the patient's blood sugar levels are measured, after remaining on fast for at least 8 hours. If a level of 126mg/dL or higher is achieved then the patient is suffering from diabetes.
2. **HbA1c test:** In this examination, the patient's average blood glucose level is recorded with a reading of 6.5 or more, denoting diabetes.
3. **Oral test:** A patient is made to drink a water mixed with high concentration of glucose and then level of blood sugar level is recorded after 2 hours. This test involves taking a sugary drink and two hours later, taking a blood glucose reading. 200 mg/dL acts as a threshold which indicates the presence of diabetes.

4. Random blood sugar test: The glucose level is recorded at any time of day with or without fasting. When combined with diabetic symptoms, a result of 200 mg/dL or above suggests the presence of diabetes.

2.2.2 Challenges and Limitations

Early diagnosis of diabetes mellitus is key to improvement of health conditions of common people and timely treatment and prevention of complications. However, there are number of challenges and limitations faced in early detection of diabetes:

1. Non-Specific Symptoms: Common symptoms of diabetes like frequent urination, fatigue, thirst are not specifically associated with diabetes but attributed to other health factors which results in delays in diagnosis and medical attention of the disease [33] [34].
2. Silent Progression: In most of the cases, diabetes progresses silently without any visible symptoms during the early stages and becomes challenging for medical practitioners to detect disease before onset [35] [36].
3. Lack of Awareness: Common man is usually unaware about common risk factors of diabetes and is not aware about importance of early diagnosis that may result in missed opportunities for early diagnosis of diabetes [37].
4. Limited access to Health Care: People living in remote areas and belonging to disadvantaged sections do not have proper access to health care which hinders early diagnosis of diabetes [38].
5. False Positives and False Negatives: Tests done for diagnosis can show false negative and false positive results [39], missing actual cases and false identification of diabetes not actually positive. Equally measuring sensitivity and specificity is very important but challenging [40].

2.3 Advancements in ML and DL

AI is a rapidly developing field that is crucial to the early identification of many diseases and the advancement of medical facilities. Diabetes, if left unchecked, can result in life-threatening health issues [29] [41]. The development of AI, particularly machine learning and deep learning, has created new avenues for identification and prediction of

DM. A few examples of how ML and DL have aided in the early diagnosis of diabetes include:

1. **Image Analysis:** Deep Learning has shown promising results in diagnosing medical images and collecting useful information from the images and using that information for early diagnosis of various diseases [24].
2. **Blood Glucose Prediction:** ML and DL can be used to read the clinical databases or data obtained from wearable devices as well as to mine the data for undiscovered patterns to predict the parameters for early diagnosis of diabetes so that the patients can be assisted in managing blood sugar levels [28].
3. **Data Fusion:** Data collected from multiple sources like medical health centers, electronic wearable devices etc., can be fed to deep learning and machine learning algorithms to study physiological characteristics like pulse rate, physical activity, resting patterns etc. [11]. Any kind of poor physical activity can trigger warning signs for diabetes and its associated complication so that the patients can take timely measures to control the diabetes [42].
4. **Real Time Monitoring:** Wearable devices equipped with sensors and ML algorithms can study data on real time basis and generate alarm signs on detection of characteristics associated with DM collected from wearable devices [29].
5. **Personalized Treatment Plans:** Various Deep learning algorithms can analyze the treatment outcomes of patients and suggest improved treatment plan that varies from individual to individual [43].

2.4 Previous Research and Models

DL models have shown good results in recent years for their capability to analyze huge volume of datasets and help in early diagnosis of diabetes mellitus. Various models proposed were applied on multiple datasets but most of the models employ PIMA INDIA dataset and Retina image dataset from early diagnosis of diabetes mellitus [44]. Deep Neural Networks for Diabetic Retinopathy identification, Recurrent Neural Networks for Glucose Level Prediction, and Deep Learning Models for Risk Prediction are a few of the most helpful models created for the early identification of DM. A number of datasets are openly accessible for the diagnosis of DM. However, the applicability of each dataset varies from region to region. Some of the most commonly used datasets for study of DM

prediction using ML and DL are:

- a) Pima India Database.
- b) National Health and Nutrition Examination Survey.
- c) Diabetes data from UCI machine learning repository.
- d) Early diabetes risk prediction dataset.
- e) Diabetes retinopathy Image dataset.
- f) Medical Information Mart for Intensive Care III.
- g) Electronic Health Record.
- h) Real world wearable data.

Among all mentioned datasets the most commonly used dataset for carrying out research on diagnosis of diabetes mellitus is PIMA dataset. This is a Standard and Benchmark dataset that contains comprehensive information required for early diagnosis of diabetes mellitus.

2.5 Literature Review

DM is a major health issue affecting a large proportion of people around the world. For effective diabetes control and to avoid complications, an early diagnosis is crucial. During last few years, DL algorithms have been gaining popularity in the area of medical diagnosis, including the early diagnosis and prediction of DM. This survey of the literature seeks to provide a summary of contemporary research using DL algorithms for early prediction DM.

2.5.1 Methodology for Literature Review

PRISMA stands for Preferred Reporting Items for Systematic Reviews and Meta-Analyses which is the compilation of data based on parameters specified in systematic reviews and meta-analyses. PRISMA offers writers a structure to guarantee clear and comprehensive reporting of their meta-analysis or systematic review, thereby, promoting critical evaluation and interpretation of the study's findings. The methodology is given in Figure [2.1](#).

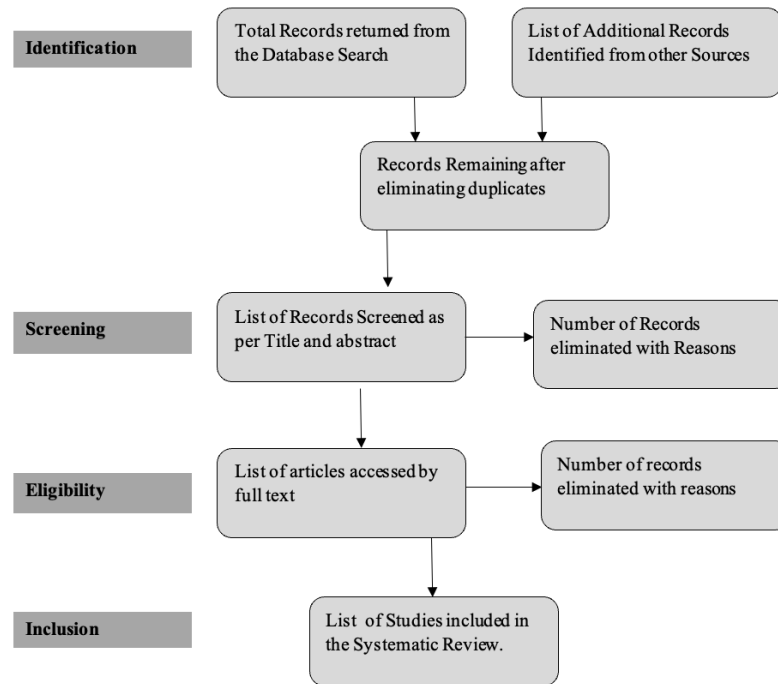


FIGURE 2.1: Methodology for Literature Review

Online resources such as PubMed, ScienceDirect, and Google Scholar were used to do a survey of the literature. The search was confined for papers published in the last five years (2018-2023). The lookup was done based on terms like "deep learning," "machine learning," "artificial intelligence," "diabetes mellitus," "early prediction," and "screening." Several studies have used deep learning algorithms for diagnosis and prediction of DM. The following are some of the significant findings of the studies: In a study by [45], the authors developed an improved model for classification of diabetes mellitus using three novel convolutional neural networks. The authors used a combination of five networks with high performance and the final network was built using CNN and LSTM. The authors conducted an evaluation of the model on 7 experiments of three datasets from Kaggle and reached an accuracy of 93.68%. A study was conducted by the writers in [46] to find the hospital-based prevalence along with the clinical feature of suddenly occurring type 1 diabetes mellitus. They collected almost 76000 records with diabetes. The authors during the study found that Type I DM was prevalent in 1.2% among total persons suffering from diabetes. The sudden diabetes known as Fulminant Type I DM was found in 3.2% people that were recently diagnosed with Type I DM. They further observed that HbA1c levels were found lower in patients that were suffering from FT1DM than the patients with non-FT1DM at diagnosis.

In a comparative study [47], the authors compared various high-end techniques of preprocessing applied in the DL-based classification of diabetes mellitus. The authors further propose an unsupervised deep learning model along with preprocessing techniques for dataset. They use K-Mean clustering along with Graham's method and Z-Score normalization technique and get the accuracy of 90.59% in classification of diabetes mellitus. A study was done by [16] to utilize NN, ML, and DL for prediction of DM. The researchers highlight the shortfalls of existing techniques that make them unrealistic for diabetes mellitus classification. They suggested the use of deep learning combined with preprocessing techniques for better results in early prediction of diabetes mellitus. An investigation including 149 individuals with Type 2 Diabetes was carried out by [12]. The authors applied DT, DNN and ensemble algorithms and designed a model. Several metrics, including accuracy, F1-Score, ROC etc were used to validate the classifier's performance. The result specifies that the ML algorithms show improved performance than DNN. The shortcoming of the study is that dataset the includes only 149 records which is not suitable for training of the predicting model.

A deep unsupervised machine learning model using voting combined feature selection and DBN for early diagnosis of diabetes was proposed by [18]. They used the dataset from Sylhet Diabetes Hospital in Bangladesh. After the application of preprocessed methods, to reduce dimensionality, features were chosen. The classifier was trained and fine-tuned utilizing feature optimization and the findings were evaluated using various criteria. The model gives the precision of 92. Researchers created a completely automated deep learning model to research diabetes CT biomarkers for diagnosis in [48]. A total of 8992 patient records were collected and the images of pancreas were segmented using a DL model. The period between the date of the CT scan and the diagnosis of type 2 diabetes was used by the investigators to categorize the patients into groups and perform univariate and multivariable analysis. The authors then determined which characteristics were most useful for type 2 diabetes prediction and trained the mathematical logistic regression model. A review study was done on Diagnosis, Risk factor and management of Gestational Diabetes by [49]. The authors used research studies from Scopus, Elsevier and PubMed and explore the effect of GDM on long term on mother and child along with health disorders that propagate to next generation. The authors explained clinical survey data and existing models for diagnosis of GDM to get an insight into the underlying pathophysiology of the disease.

In a research conducted to examine the significance of laboratory diagnosis and the limitations of conventional approaches in the diagnosis of DM by [50]. The authors studied the effectiveness of Fasting Plasma glucose test, Hemoglobin A1C or HbA1c test, random

and oral glucose test for the timely diagnosis of DM. They further suggested comparing the effectiveness of new diagnostic tests like Glycated Albumin against traditional laboratory tests and to find new techniques and tests in future for better diagnosis of the disease. A research study was done by [51] to evaluate the effect of cinnamon in controlling glucose level in blood of patients suffering from Type II diabetes. The authors designed a Quasi experimental pretest post-test control group and carried out the experiment to see the effect of cinnamon on 40 patients suffering from diabetes mellitus. From the study it was found that cinnamon helps people with type 2 DM reduce their blood glucose levels significantly and should be used in helping insulin perform its task of controlling the blood sugar level.

E-Healthcare systems face challenges that can be addressed through the deployment of Internet of Things (IoT) apps using Cloud Computing (CC) approaches. However, some drawbacks to using CC still persist, including response time, transfer rates, energy consumption, and security and privacy concerns. To overcome these obstacles, fog computing (FC) has been proposed as a CC development. A system called DiaFog has been proposed by researchers in [52] for actual and accurate diagnosis of DM disease (DMD) based on fog computing, cloud computing and Internet of Things, along with an ensemble DL (EDL) technique. The PIDD and the HFGDD, which were retrieved from the UCI-ML and Kaggle repositories, respectively, were the two datasets used to train the system related to diabetes mellitus illness. A number of tests have been conducted on the proposed system, including recall, F-measure, accuracy, precision etc. Patient diagnosis of diabetes at a distance was made possible by the integration of IoT, fog, and cloud. The trial findings show how well FC principles work when used and how well they work when used for prompt remote diagnosis of diabetes patients.

A common symptom of diabetes is the presence of a foot ulcer, which, if not detected quickly, can result in major problems and possibly amputation. The authors in [53] focused on the early detection of diabetic foot ulcers to prevent such outcomes. Infrared imaging is a suitable tool for gathering information to diagnose various diseases, including foot ulcers. This non-invasive method is faster than traditional imaging techniques. The study analyzes temperature variations in the feet of over 60 people (61.67% male and 38.33% female) and clearly classifies the risk of foot ulcers. This method is more easily understood than deep learning techniques and simpler to implement. It uses infrared imaging to give a non-invasive method of diagnosing diabetes patients' foot ulcers.

A study regarding use of traditional laboratory tests for diagnosis of diabetes mellitus was carried out by [54]. Correct diagnosis of diabetes and pre-diabetes necessitates the FPG test and A1C test, although the RPG test may be appropriate for some instances.

Recent studies have demonstrated excellent laboratory proficiency in determining fasting, HbA1c, and OGTT glucose levels. Nonetheless, in situations where more precise glycemic status is needed, traditional diagnostic tests can be compared with latest diagnostic tests, including glycated albumin (GA). The authors expect that upcoming laboratory diagnostic techniques will provide tremendous benefits for therapy efficiency and survival. Consequently, there is an important requirement to explore more tests or to substitute the HbA1c and OGTT with more current laboratory diagnostic methods for diabetes mellitus. A research was carried out on how diabetes results in heart health issues and how HRV signals can indicate the presence and severity of diabetes by detecting diabetes-induced cardiac impairments by [55]. Analyzing HRV signals, which are non-stationary and nonlinear, can be extremely difficult, but DL methods, have been shown to effectively extract useful information and identify correlations between diabetes and HRV signal variations quickly and accurately. The Authors also study several DL architectures that can be utilized to analyze HRV signals and detect diabetes. Deep learning methods are currently the most advanced techniques for analyzing HRV signals and detecting subtle changes from normal. Deep learning networks can be scaled up to process huge amounts of data in a scattered manner, and distributed DL algorithms can be used to learn patterns and make correct diagnosis about the future progression of the disease.

In a study by [18], the authors offer an Unsupervised deep learning model using DBN and voting ensemble feature selection for prediction of DM. The dataset, which included replies from pre-diagnosed patients who filled out several questionnaires made at the Bangladeshi Sylhet Diabetes Hospital, was retrieved online. An ensemble feature selector was used to apply the preprocessing and feature reduction to the dataset. After training and making adjustments to reach peak performance, the DBN model was contrasted with a number of alternative models that lacked different hidden layers. With an F1-measure, accuracy, and recall of 1.00, 0.92, and 1.00, respectively, the DBN model performs comparatively well. The study comes to the conclusion that DBN is a useful technique for the unsupervised early detection of Type II DM based on these findings. The researchers in [56] gathered clinical samples from 1000 pregnant women, including 221 cases of GDM. The imputation of missing values was done by utilizing matrix factorization approach. Then RF model was used to identify important clinical attributes to diagnose GDM by assessing the significance of every feature dimension. Ultimately, TF-GDM, a novel transformer-based method, was created that accurately aids in GDM prediction. The outputs demonstrate that the TF-GDM technique works well instead of the conventional ML and DL based approaches, with improved recall, accuracy, and precision rates of 0.92, 0.88, and 0.93, respectively, as well as an F1 Score and an AUC

value of 0.90 and 0.94. The research carried out by [57] collected data for 52,139 in persons suffering with T2DM from year 2008 go 2016. These notes were processed using a transformer architecture-based symptom annotation model called T5-depression, which helped identify depressive symptoms from the patient's present illness. The F1 score and the AUROC were utilized to evaluate the efficacy of the model. They also analyzed the connectivity of depressive symptom networks in T2DM patients, including those with complications. In a research by [58], the authors studied the association between Pancreatic neoplasia and diabetes mellitus and observed to have an association, and this relationship has been the subject of their research. They found that Early detection and screening for pancreatic neoplasia are essential for improving patient outcomes. However, pancreatic neoplasia is challenging to detect since there are no particular symptoms and the screening instruments are non-invasive.

In a research study by [8], the authors focused on a combined deep learning approach that incorporates both LSTM and CNN architectures to categorize DNA sequences for the gene of insulin and forecast the potential for diabetes due to changes in the gene sequence. Several performance indicators, including accuracy, precision etc, were used to evaluate the efficacy of the proposed model. The outcomes of the experiment demonstrate that the suggested model produced the best outcomes. The hybrid LSTM-CNN model demonstrated an accuracy of 99% during the learning phase, whereas the CNN and LSTM models had accuracies of 97.5% and 95% respectively. The authors in [59] studied the preprocessing techniques and their role in the success of ML models, as they worked to improve the input data quality by addressing various problems with data quality, such as noise, missing values, and irrelevant attributes. They proposed a technique that combines missing value imputation and feature selection methods to enhance the employed classifier's performance on a reputable DM dataset.

Based on the study by [60], the Freestyle Libre Pro 2 continuous glucose monitoring device was more acceptable to pregnant women as a tool for diagnostic test for GDM than the OGTT. The study also used a combination of CGM parameters, OGTT results, and GDM risk factors to create a Total Risk Score (TRS) and a CGMSV to triangulate the results. They suggest that CGM devices such as Freestyle Libre Pro 2 could be a more thorough and palatable substitute for OGTT in the diagnosis of GDM. Combining CGM data with additional risk variables and ultrasonography characteristics might increase precision and decrease false positives and negatives.

2.6 Complexities with Design and Development

Based on the literature review, various complexities that are associated with Design and Development of a Model for Diagnosis of Diabetes Mellitus have surfaced as well as the techniques used to handle those complexities. The remaining part of the section focuses on these complexities.

2.6.1 Challenges in Diabetes Diagnosis

In recent times, diabetes, which is among the most prevalent diseases globally, has emerged as a growing danger to human health on a global scale. On the other hand, diabetes is significantly slowed down in its progression when detected early. For the disease to be stopped from spreading, early detection is crucial. The growth can only be stopped by early identification of the disease because diabetes is a lifelong condition that has no known cure [61]. On the other hand, a late diagnosis could lead to cardiac problems and major organ damage. Prediction of diabetes is often aided by the use of both clinical and physical data, such as serum insulin, BMI, age, and plasma glucose levels [11]. This data indicates that a physician makes the disease diagnosis, but diagnosing a patient is a highly challenging and time-consuming procedure for the physician. Furthermore, the doctor's judgments could be biased and incorrect. Due to this, the domains of data mining and ML are commonly used as a DSS for the quick and precise identification of illnesses based on data [62].

2.6.2 Diagnosis using Artificial Intelligence

Algorithms that let computers complete human tasks more quickly and automatically have recently arose from the development of computer technologies. Artificial intelligence techniques like ML and DL have demonstrated impressive results in analyzing current data [63]. Artificial intelligence-based techniques are particularly useful in the medical industry for the quick and effective diagnosis and treatment of a wide range of illnesses. Cancer, diabetes, COVID-19, heart illnesses, brain tumors, Alzheimer's, and other diagnostic investigations are a few examples of these. The medical industry can benefit greatly from artificial intelligence. Big data in medicine has lately become common place in hospitals due to artificial intelligence's greater performance in research projects. Given that every patient is a unique data point, a significant amount of numerical data, including ECG, EMG and a multitude of other data, including computed

tomography (CT), MRIs, and X-rays, can be generated following the review of medical data [64]. In this sense, a sizable amount of big data is composed of these medical records.

Big data based on AI is typically interpreted (by regression, classification, or clustering) using ML techniques. These algorithms enable the identification of the interrelationship between them using observations of data and samples. AANN, SVM, k-NN, DT, and NB are some of the ML techniques that are commonly utilized in this field. The association between the independent and target data is directly learned by these methods [65]. But during the past ten years, advances in computer processing power and artificial intelligence have deepened ANN, leading to the rise of DL—which combines feature extraction and categorization [66]. DL has provided a leading edge over ML algorithms, especially in big data applications. Convolutional neural networks (CNNs) are the most often utilized model in DL-based clinical diagnosis and detection applications [25]. Because of its deep architecture and advanced feature representation, CNN models are highly used. Since CNN was built with an end-to-end architecture, classes are produced as output and raw data are provided as input. Consequently, the CNN model's performance is significantly impacted by the architecture's design [48]. But recently, researchers have started using famous CNN designs like ResNet, GoogleNet, Inception, Xception, VGGNet, and others, together with transfer learning applications. In many data-driven research projects, there are benefits to using pre-trained or pre-designed CNN architectures directly, including ease of use and improved performance [8].

2.6.3 Deep Learning and Numerical Dataset

The dataset is the primary determinant of the model's performance. If the dataset is adequately large to train the model, the model will make more accurate and unbiased classification [41]. However, if the size of the dataset is small, it may result in development of a model which will make unreliable classification. From the literature survey it was found that most of the research done for early diagnosis of DM was carried out by ML or DL algorithms using PIMA dataset [67]. The dataset is suitable for machine learning model training due to its small size. The PIMA dataset is tiny, and this issue has to be fixed in order to take advantage of deep learning's advantages for managing a chronic illness like diabetes mellitus [22]. Besides most of the commonly used deep Neural Networks works on images, this also makes application of Numeric dataset like PIMA unsuitable for designing a DNN model for early diagnosis and detection of diabetes.

Numerous medical data in the clinical sector are made up of numerical values, just like

the PIMA dataset. It is more typical to use numerical values directly using "conventional machine learning techniques", which is evident in studies associated with ML models like "SVM, NB, RF, DT" [9] [68] [69] [70]. The PIMA characteristics are fed to the fully linked layers or the 1D convolution layer in studies that generate deep learning models using the same data. "Recurrent neural network (RNN)-based long short-term memory (LSTM)" was utilized in some studies to handle the PIMA dataset, which contained 1D data [71]. However, the PIMA dataset has independent data, whereas LSTM was created for sequential data.

Deep learning, which has gained popularity recently, provides number of advantages over conventional machine learning algorithms with the elevated level of features They provide deep "CNN models" in particular, which have demonstrated superior efficiency. However, up until now, researchers have created 1D CNN models based on the numerical values from the PIMA dataset. Given that widely used CNN models require only two-dimensional data to be entered into the input layer. Applications for transfer learning use these models [72].

Consequently, using the PIMA dataset comprising autonomous numerical data, feature extraction by widely used CNN models and a diabetes prediction that makes use of these models are still in development. Therefore, in order to provide diagnoses that are more accurate, the raw data may be transformed in accordance with commonly used CNN models [72].

2.7 Comparative Analysis

Increased quantities of glucose cause a metabolic disorder called diabetes mellitus (DM). If ignored, DM can cause a number of diseases related to the heart, liver, and brain. The global high mortality rate caused by DM has caused havoc worldwide. However, with the increasing use of ML and DL algorithms in making predictions in eCommerce and better business decisions, there is a ray of hope for using these techniques in medical science to assist in the timely prediction of various diseases. Today, with a vast volume of medical data available, there is a possibility to apply ML techniques to these datasets and find useful patterns and hidden information that can later be used to predict diseases much earlier before their onset.

In machine learning, classification involves building a model that identifies and categorizes a dataset into distinct classes, while clustering is a process that examines data objects without utilizing class labels, grouping samples into new classes by maximizing the similarity between them. Association Rule Learning (ARL) is another approach that

mines frequent patterns from data.

In 2018, around 11% of the United States population was affected by diabetes, with 1/5 of those cases being undiagnosed. Unfortunately, many individuals are unaware of their susceptibility to diabetes until the disease has progressed significantly. Therefore, early detection of diabetes is essential to avoid severe problems. While diabetes cannot be cured completely, early detection can aid in reversing some of its effects and help patients achieve remission by maintaining normal blood sugar levels without long-term medication. ML and DL may have a significant impact in early detection. The medical industry generates vast amounts of data from hospitals, nursing homes, clinical health centres, and polyclinics, making it challenging to process manually. Employing various DL algorithms can extract hidden relations and information from the datasets and forecast the onset of diabetes before the disease progresses. This proactive approach enables necessary measures to be taken to prevent multiple health-related problems in patients and help them lead healthy and fulfilling life. However, the raw dataset may contain multiple anomalies, such as missing values, redundant information, null values for some attributes, and erroneous values, making it challenging to apply DL algorithms to it. Therefore, the dataset must be processed and converted into a usable form that aids informed decision-making.

2.7.0.1 Machine and Deep Learning

With technological advancements, the lifestyle of modern individuals has become increasingly comfortable, leading to a reduction in physical activity and a rise in various health issues, including diabetes mellitus (DM), which has become a significant problem in the last two decades. The diagnosis and efficient treatment are challenging due to its complex mechanisms and related symptoms. AI is everywhere, especially in healthcare, including the use of ML and DL to process large datasets generated by medical industries such as hospitals, nursing homes, and clinical laboratories. ML and DL algorithms can extract hidden patterns and information from these datasets, which are too large to be processed manually. By applying ML algorithms to diabetes datasets, researchers can predict the onset of diabetes and potentially enhance the health outcomes of the world population. The number of the most relatable paper on the PIMA dataset using ML and DL techniques is given in Figure 2.2, these include the papers containing topics involving machine learning based methods for the diagnosis of Diabetes Mellitus and Figure 2.3 encompassing the publications containing papers with studies on deep learning-based methods for the diagnosis of Diabetes Mellitus respectively.

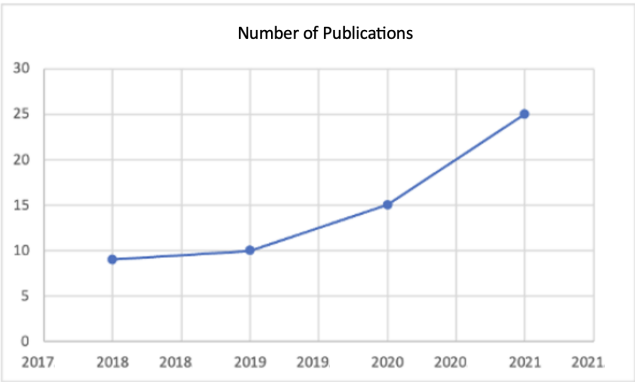


FIGURE 2.2: ML Based Publications, ML techniques for Diabetes Diagnosis

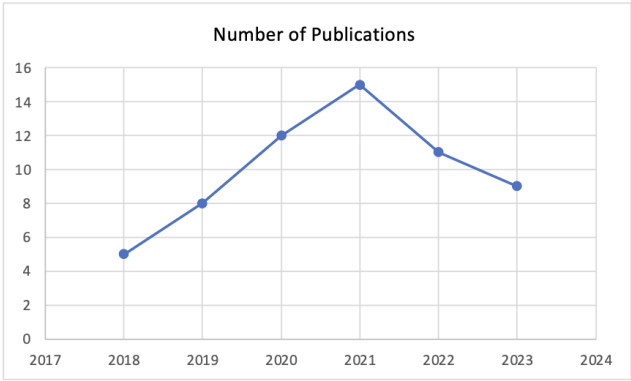


FIGURE 2.3: DL Based Publications, DL techniques for Diabetes Diagnosis

Table 2.1 and Table 2.2 summarize some of the ML and DL-based research and comparative analysis conducted in this field, respectively.

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method			Dataset	Best performance	Perfor- mance Method	Result of Best Performance Method
[73]	2021	Various	ML	Tech- niques	Local Dataset	SVM		Accuracy: 98%
[44]	2021	Various	ML	Tech- niques	PIMA	KNN, Regression	Logistic	Accuracy: 80%

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[74]	2021	LR, ANN, DT, NB DNN BayesNet, AdaBoost, Decision Bagging, RF, Proposed Ensemble Model	PIMA	Proposed Ensem- ble Model	Accuracy: 79.22%
[75]	2021	J48, CART and Naive Bayes SVM Logistic Red. Logistic Step Elastic Net LGBM: BstLinTree LDA XGB: Tree	Local Dataset	J48 and CART	Accuracy: 99%
[76]	2021	LGBM: Boost Tree XGB: Linear C5.0 Rand F. Red. LGBM: RF CART Naive Bayes Red. K/TF DenseNN	PIMA	LGBM: Boost Tree	Accuracy: 93.44%
[77]	2021	Various ML Tech- niques	Local Dataset	SVM	Accuracy 97.87%
[78]	2021	LR, LDA, NB, K- NN, CART, SVM	PIMA	Naive Bayes	Accuracy 95%

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[79]	2021	Naive Bayes	Early-stage diabetes risk prediction dataset	Random Forest	Accuracy 99.3%
		Neural Network			
		AdaBoost			
		kNN			
		Random SVM			
[80]	2018	Back propagation	PIMA	Back Propagation	Accuracy 83.11%
		J48			
[81]	2021	NB, SVM	PIMA	Ensemble Method	Accuracy 87.09%
		RF, LR, DT, SVM, NB, KNN, EM			
[82]	2021	DT	PIMA	Decision Tree	Accuracy: 71.35%
[83]	2018	DT, KNN, SVM, RF, NB, LR	PIMA	Random Forest Linear Regression	Accuracy: 90%
		RF			
		KNN			
		MLP			
		Ada boost			
[84]	2022	D tree Classifier NB	Local Dataset	Stacked ensemble with genetic algorithms	Accuracy: 98%
		GBC			
		SVM			
		Extra Tree Suggest Method (ST-GA)			

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Perfor- mance Method	Result of Best Performance Method
[85]	2021	Radial Basis Neural Network Function Genetic Algorithm DUNN Index Davies Bouldin Index Silhouette Index	PIMA	Logistic Regres- sion		Accuracy: 80%
		DT				
		LR				
		SVM				
		KNN				
		NB				
		GB				
[86]	2021	LR, KNN, SVM, NB, DT, RF, Soft Voting Classifier, AdaBoost, Bag- ging, GradientBoost, XGBoost, CatBoost ANN	PIMA	Soft Voting Clas- sifier		Accuracy:79.08% Precision: 73.13% F1 Score:71.56% Recall:70%
[87]	2019	Random Forest Clustering SVM	PIMA	ANN		Accuracy: 75.8%
[88]	2019	RF CNN	PIMA	RF		Accuracy: 83.67%
[89]	2019	SVM RF, NB, DT, KNN	PIMA	SVM		Accuracy: 77.73%

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Perfor- mance Method	Result of Best Performance Method
[90]	2019	J48	PIMA	Logistic Regression	Regres- sion	Accuracy: 77%
		NB RF, LR,				Precision: 0.77 Recall:0.77 F-Score:0.76 AUC: 0.83
[91]	2021	ADA BOOST with RF	Image Dataset taken from local Clinic	Ada Boost with RF		Accuracy: 96.71%
		ADA BOOST with Extra Tree				Precision: 97.55 Sensitivity: 97.95 F1-Score: 97.75
[11]	2022	KNN CNN	PIMA	Extreme Learning		Accuracy: 90.54%
		SVM LR Extreme Learning				
[92]	2018	NB SVM	PIMA	Naive Bayes		Accuracy: 76.3% F-Measure: 0.76 Precision: 0.759 Recall:0.763
		DT				
[93]	2020	LR KNN	PIMA	Random Forest		Accuracy: 75.0% Sensitivity: 0.250 Specificity:0.789 Precision:0.661
		SVM NB DT RF				

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[94]	2020	RF	Local Dataset	RF	Accuracy 99.35%
		SVM			SEN 99.01 %
		AdaBoost			SPE 100%
		Gradient Boosting			FPR 0%
					FNR 0.99%
[32]	2021	LR	PIMA	KNN	NPV 98.15%
		KNN			Precision: 0.747
		SVM			Recall: 0.751
		NB			F-Measure: 0.749
		DT			Accuracy: 75.10%
[95]	2020	RF	PIMA	Proposed CNN	
		KNN			
		LR			
		DT			Accuracy:
		RF, SVM			93.2%
[96]	2022	MLP classifier	PIMA	Linear SVM	
		Proposed CNN			
		Linear Kernel SVM			Accuracy: 89%
		Radial Basis Kernel			Precision: 0.87
		SVM			Recall: 0.88
		KNN		Kernel	F1-Score 0.87
		ANN			AUC: 0.90
		MDR			

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[97]	2021	RF (cross-validation)	Local Dataset	Random forest (cross-validation)	F-measure: 0.983 MCC: 0.9654 AOC RUC: 0.999 PR AUC: 0.999 Accuracy: 98.3055%
		NB (cross-validation)			
		KNN (cross-validation)			
		J48(cross-validation)			
		RF (split method)			
		NB (split method)			
		KNN (split method)			
[98]	2021	J48 (split metho	PIMA	decision tree	Accuracy: 85%
		NB			
		DT			
		SVM			
[99]	2018	SVM	PIMA	Proposed Method (PM)	Accuracy: 90.36%
		Bayes Net			
		DecisionStumb			
		AdaBoostM1			
[100]	2021	Proposed method	Pregnant cohort study in eastern China	Random Forest	Accuracy: 86.91% Sensitivity: 63.30 Specificity: 97.53 AUC: 0.80
		LR			
		RF			
		SVM			
		ANN			
[101]	2021	ANN	PIMA	Logistic regression	Accuracy:77.61% Recall:0.8902 Precision:0.7979
		SVM			
		K-NN			
		DT			
[102]	2021	NB	Local Dataset	DLCNN	Accuracy: 98.42%
		LR			
		DLCNN, CTCNP,			
		LVQOAC, MODLNN			

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Perfor- mance Method	Result of Best Performance Method
[103]	2020	KNN	Wisconsin dataset	Decision and Regression	Tree and Logistic	Accuracy: 97%
		NB				
		RF				
		SVM				
		DT				
[4]	2021	LR	HbA1c- labeled and FPG- labelled datasets	SVM		Accuracy: 82.10% Precision: 82.30 Recall: 82.10 F1 Score: 82.05
		LR				
		SVM				
		DT				
		RF				
[104]	2018	SVM	Messidor	SVM		Accuracy: 90.04%
[105]	2018	RT	Chronic Kidney Disease Dataset from Apollo Hospital	Logistic gression Multilayer ceptron	Re- and Per-	Accuracy:98.1% F1 score:98.4
		SVM				
		LR				
		MLP				
[106]	2018	RF	PIMA	MLP neural net- work		Accuracy: 77.08%
		LR				
		MLP neural network				
[107]	2021	Ensemble of ADA Boot	PIMA	Ensemble of ADA Boot		Accuracy: 95.0%
		XG Boost				

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[108]	2021	RF	PIMA	Random Forest	Accuracy: 94.0%
		DT			
		NB			
		LR			
		ADA Boost			
[109]	2019	SVM	Diagnostic dataset from medical Center	C4.5 Tree	Accuracy: 74.0%
		NB			
		KNN			
		C4.5 DT			
[110]	2020	KNN	PIMA	Random Forest	Accuracy: 74.47%
		SVM			Precision: 80.48
		RF			Recall: 79.83
					F1-Score: 80.16
[111]	2020	K-Means Algorithm	PIMA	SVM	Accuracy: 93%
		LR			
		SVM			
		KNN			
		RF			
[112]	2018	DT	PIMA	SVM	Accuracy: 79.13%
		NB			
		NB			
		SVM			
		RF			
[113]	2018	Simple CART	PIMA	SVM and KNN	Accuracy: 77%
		SVM			
		KNN			
		LR			
		DT			
		RF			
		NB			

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[114]	2019	RF	UCI Learning Repository	Random Forest	Accuracy: 90%
[115]	2019	RF	Clinical Dataset	Random Forest	Accuracy: 95.1%
[116]	2020	Glmnet RF XGBoost LightGBM)	Clinical Dataset	Glmnet	Accuracy: 95%
[117]	2021	Various ML Techniques	Diabetes Hospital of Sylhet	RF	Accuracy: 99%
[118]	2020	RF XGBoost	PIMA	XGBoost	Accuracy: 74.10% Precision:0.701 Recall: 0.817
[119]	2019	Linear Discriminant Analysis (LDA)	PIMA	LDA	Specificity: 0.720 F-Score: 0.755 Accuracy: 76.86%
[120]	2021	ANN NB DT SVM	Data collected from android application and PIMA Dataset	SVM	Accuracy:81.6% Sensitivity:87.32 Specificity:73.46

TABLE 2.1: Comparative Analysis ML-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[121]	2021	RBF	PIMA	RBF	TP Rate: 0.459 FP Rate: 0.819 Precision:0.792 Recall:0.860 F-Measure: 0.825 MCC:0.459 Recall: 0.792 ROC Area: 0.890 TPR: 77.36 TNR: 89.11 FPR: 10.89 FNR: 22.64
[122]	2021	KNN	PIMA	K-Nearest Neighbor	F1 score:78.10% Accuracy:85.06% Recall:77.36 % Precision:78.85% Specificity:89.11% Sensitivity: 99.56
[1]	2021	DT AdaBoost RF	PIMA	Random Forest	Positive predictive value: 93.25 Negative predictive value: 89.98 F-measure: 96.30
[123]	2021	SVM XG Boost	PIMA	XG Boost	Accuracy: 77.0%

TABLE 2.2: Comparative Analysis DL-based Techniques

Ref.	Year	Method	Dataset	Best Performance Method	Result of Best Performance Method
[124]	2022	Deep Neural Net-work	PIMA	Deep Neural Net-work with missing values handling	Accuracy: 80.0 (MAX)
[125]	2020	Deep Learning Decision Tree Artificial Neural Network Naïve Bayes	PIMA	Deep Learning	Accuracy: 91.07
[126]	2021	Deep Learning SVM	PIMA	Deep Neural Net-work	Accuracy: 77.474
[127]	2019	Logistic Regression Improved GA Modified K-Means + SVM SVM with efficient coding Deep Neural Network	PIMA	Deep Neural Net-work	Accuracy: 89.35
[122]	2021	Deep Learning TLSTM CLSTM	PIMA	Deep learning	Accuracy: 93.7% Accuracy: 91.6%
[128]	2022	DNN + 10-fold cross-validation	PIMA	Deep Neural Net-work	Sensitivity: 87% Specificity: 91% Accuracy: 89%

Many research studies do not provide a single classification model for predicting the two classes of diabetes. There has been the use of a single dataset with few records which doesn't provide reliable results. Many researchers use the PIMA dataset without any pre-processing technique like normalization. As a result, the results suffer from outliers, overfitting, underfitting and other anomalies. There have been studies that used limited machine learning algorithms to diagnose diabetes and did not handle the missing values. In other studies, some authors have inhibited the application of feature extraction fully. The feature extraction process could be enhanced by the application of an automatic

process of deep feature extraction.

Some studies do not consider the importance of all the attributes of the dataset. Attributes like body size, height and BMI and their role in the DM diagnosis have not been considered, which affects the performance of the classifier. Many authors investigated only matricellular proteins as biomarkers however there are multiple biomarkers like microRNAs, angiographic vasospasm etc. Some models suffer from the anomaly of oversampling. It has also been brought to light that the medication affects the attributes of the patients, many researchers in their research did not collect any data regarding the medication of patients which limits the efficiency of the classifier. The limitations of various research studies in diabetes prediction are enlisted in Table 2.3 as under:

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[73]	The study focus on traditional datasets and did not explore the capability of deep learning architectures fully. The study also lacks standard evaluation metrics, has many computational constraints and interpretability problems. It is essential to address these limitations in order to develop more accurate and reliable artificial intelligence based model for the early diagnosis of diabetes.
[44]	The authors highlight the benefits of accurate diabetes prediction on time and compares various machine learning models. The proposed IDMPF model shows better results. However, it has the limitation of poor data driven performance. Besides, the study focuses on two models and no work has been done towards class balancing. More models need to be explored with bigger datasets and balancing techniques in order to improve the accuracy and generalizability of the model.
[74]	In this study, an ensemble model that integrates J48, NBTree, Random Forest and Simple Cart machine learning techniques for early diagnosis of Type 2 diabetes. The limitation of the study is that it relies on single normalization technique. Besides, the integration of machine learning and deep learning techniques could improve the accuracy of model which was not explored in this study.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[75]	Various data mining and machine learning models have been studied and compared. SVM and NB give promising results. The study has limited scope of algorithms, depends on dataset and does not apply feature engineering for feature importance and feature selection. Working on these limitations could result in development of more comprehensive and robust diabetes prediction model.
[76]	The study highlights the limitations in multilayer perceptron model that can be addressed by data balancing techniques to improve the performance of the model.
[77]	The authors have worked on comprehensive review of various ML techniques for early diagnosis of diabetes along with focus on importance of data preprocessing techniques like feature selection, data denoising and feature extraction to improve the accuracy. The limitations pertain that only basic machine learning models have been used for comparison without considering deep learning approaches. Therefore, a clear and more complex picture of the state of the art models for diabetes prediction is not covered in the study.
[129]	This study focus on comprehensive review of data mining techniques for the early diagnosis of various endocrinal diseases with main focus on diabetes mellitus and thyroid disorders. The study did not focus on classifiers that works on fewer features for fast diagnosis.
[79]	The study uses various data mining techniques for early diagnosis of diabetes mellitus using limited feature dataset that does not have features like family history, prescription drugs, smoking and sleep deprivation. These features can improve model accuracy and generalizability
[80]	The study shows that backpropagation achieves 83 percent of accuracy that over performs other machine learning algorithms for diagnosis of diabetes mellitus. The limitation of the study is that it implements limited number of algorithms with little iterations. More advanced machine learning algorithms and deep learning algorithms with more epochs can be used to give more clear picture of the performance of various algorithms for early diagnosis of diabetes.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[130]	The authors highlight the advantages of AI based tools for early prediction of diabetes mellitus. The study shows the better performance of ensemble techniques over individual machine learning algorithms. The limitation of the study is that it considers various studies for review that are more than decade old. More recent approaches can be explored with more evaluation metrics for better results.
[82]	While applying decision tree for early prediction of diabetes mellitus, the authors concludes that 50 percent split of dataset into training and testing provides best accuracy. However, this is neither true for all datasets nor for all algorithms. The study only shows the performance of decision tree without considering other state of art machine learning algorithms for diagnosis of diabetes mellitus.
[83]	The authors implement modified J48 classifier for early prediction of diabetes mellitus along with other classifiers. The authors compare various algorithms based on the accuracy, however accuracy is not always the best evaluation metrics and may lead to biased evaluation in case of imbalanced datasets. Considering more advanced evaluation metrics could provide better picture of the study.
[84]	The study uses PIMA datasets for early diagnosis of diabetes. The proposed algorithm achieves high accuracy, however, the applications of ensemble algorithms and deep learning algorithms have not been considered. Addressing such limitations could improve the generalizability and robustness of proposed approach.
[85]	The study proposes the development of model for diabetes diagnosis using cluster validity index, k-means clustering and Radial Basis Neural Network. The limitation of the model is that it considers only adult population while predicting diabetes mellitus. Besides the comparison of proposed model with limited number of models also does not give clear picture of the robustness and generalizability of the model.
[86]	The authors propose a soft voting classifier model that is an ensemble of RF, LR and NB for early diagnosis of diabetes mellitus. The limitation of the study is that they do not consider the application of deep learning techniques for diagnosis of the diabetes mellitus.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[87]	The research stresses on the efficiency of machine learning algorithms in early prediction of diabetes mellitus with ANN outperforming other algorithms, however, the limitation of study is that they carry out the research on single structured dataset with limited number of records and no features for taking relevant features into account like smoking, hereditary traits etc.
[88]	The authors compare various machine learning and deep learning algorithms and find Random Forest is the most effective in predicting diabetes mellitus. The limitation of the study is that the use of feature extraction using automatic deep feature extraction approaches of deep learning have not been considered. Future research could emphasize on application of better feature extraction techniques for improving the performance of the model.
[89]	A novel approach for early prediction of diabetes mellitus was proposed by the authors using simple feature selection techniques. There is a scope of improvement by exploration of advanced feature engineering techniques and with analysis of correlation of various features with the target variable.
[90]	The review study of various machine learning algorithms for diabetes prediction shows Logistic regression to be most effective for diagnosis of diabetes mellitus. It only focuses on traditional machine learning algorithms without taking unsupervised machine learning algorithms and deep learning techniques into account.
[91]	Diabetic foot ulcers are one of the diseases caused due to diabetes. The study focuses on the development of a diagnostic model for early detection of foot ulcers using 2D image based deep learning models. The limitation of the study lies in its dependence on images of particular population that compromises the models generalization.
[11]	A review article for early detection of diabetes using machine learning algorithms was presented in this study hat lacks taking into consideration most recent papers and focuses on study of research carried between 2015 and 2020 thus overlooking most relevant recent studies.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[92]	The study proposes a machine learning approach for early prediction of diabetes mellitus in which Naïve Bayes outperforms other models. The downside of the study is that only few basic machine learning algorithms were taken into account without incorporating advanced machine and deep learning techniques that may give more better performance in prediction of the disease.
[93]	The study focuses on prediction of Type 2 Diabetes using basic machine learning algorithms. The limitation of the study is that the algorithms were trained on small size dataset. Besides, advanced machine learning techniques were not taken into account for the classification of disease that may improve the outcome of the study.
[94]	The authors emphasize the application of preprocessing techniques on dataset used for early prediction of diabetes mellitus. The downside of the study is that advanced preprocessing techniques were overlooked and basic preprocessing techniques were used that may have the potential to introduce bias in the results.
[32]	The study has been done for early prediction of diabetes mellitus using various data mining techniques in which neural network outperform other prediction models. The downside of the study is dependence on basic machine learning algorithms with considering advanced deep learning models. Future research could explore more advanced ML and DL algorithms to improve the accuracy of the prediction model.
[95]	The study focuses on early diagnosis of diseases using machine learning and deep learning. The CNN algorithms outperforms various other model is disease diagnosis. It takes into account general diseases without any focus on diabetes dataset.
[96]	The authors in this study propose five ML models for early diagnosis of diabetes mellitus. They further use basic feature selection to reduce the dimensionality of the dataset. Taking into account advanced feature selection techniques along with missing value imputation may enhance the outcome of study. Further application of deep learning models can also lead to better results.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[131]	The authors focus on contribution of machine learning models for diagnosis of diabetes mellitus. The authors did not take into account important features like BMI, Height, and Weight while prediction of the disease that could lead to poor prediction of the model.
[98]	This review study highlights the effectiveness of various machine learning algorithms in prediction of diseases like cancer, diabetes and tumors. It focuses on limited machine learning algorithms without consideration of deep learning and advanced unsupervised machine learning algorithms.
[99]	In this study, a detailed review of various data mining techniques for diabetes prediction was given showing ensemble methods outperforming individual methods. The limitations of study is dependence on small dataset without any data augmentation techniques that may result in over fitting of the models.
[132]	This study presents an effective comparison of different machine learning algorithms for the diagnosis of Gestational Diabetes Mellitus and proposes a new stacking model that shows better performance than other models. The limitation of the study is that they compares the results with only limited set of algorithms and the experiment was carried out without any clinical setting. Considering more models in future can enhance the performance with better results.
[101]	This review article presents various data mining techniques used for early diagnosis of diabetes and also highlights the effectiveness of various techniques for diagnosis of diabetes and early preventive measures. The limitation of the study is that there is focus on particular techniques without taking into consideration advanced machine learning techniques like unsupervised techniques. Future research could address such limitation and considering unsupervised and deep learning techniques to give more comprehensive understanding and better results.
[102]	The study combines Deep Learning Convolution Neural Networks with Harris Hawks Optimization (HHO). The proposed model gives high performance as compared to traditional machine learning models but is limited by its dependence on text based dataset that is not suitable for deep learning models like CNN. Besides, the number of records in dataset are very few that may introduce over fitting in deep neural networks.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[103]	A comprehensive review of various machine learning algorithms is given in the study for diagnosis of breast cancer and Kidney disease. The study considers only few machine learning algorithms like SVN, KNN, RF and NB. A more comprehensive study with focus on unsupervised and deep learning models is needed to clear picture of the field.
[133]	The comprehensive study of various machine learning techniques is given in the study for early prediction of diabetes mellitus. Various machine learning algorithms were considered among which HbA1C and FPG gives better performance. The downside of the study is no consideration of deep learning approaches for the prediction and missing of preprocessing techniques.
[104]	The study focuses on contribution of machine learning techniques for early diagnosis of diabetes and highlights various areas where improvements are required for better prediction of the disease. The study lacks generalization and does not take into account real world factors during the study like lifestyle, economic conditions and hereditary traits.
[105]	The study was carried out to review various machine learning algorithms for early diagnosis of Chronic Kidney Diseases by considering Logistic Regression, SVM, Multilayer Perceptron and Recursive Partitioning. The LR and MLP gives better results over other ML algorithms however the downside of the algorithm is that it does not take into account other machine learning algorithms and does not focus on feature engineering techniques.
[106]	Two different models are proposed by authors in this study for early prediction of diabetes and compare the effectiveness in terms of accuracy and computational requirements. The data-recovery integrated with neural network approach offer better performance for diagnosis of diabetes. The downside of the study is application of basic machine learning techniques with much focus on preprocessing of datasets.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[107]	The study focuses on importance of correct prediction of diabetes and the capability of machine learning techniques towards achieving this goal. The downside of the study is lack of proper specification and accuracy of the proposed model. Future contribution would be to develop novel model with focus on individual patient characteristics and incorporation of real world data for study.
[108]	The study proposes a novel model for early stage risk prediction of Kidney diseases with data obtained from wearable devices and machine learning algorithms. The downside of the study is dependence on few machine learning algorithms without considering deep learning techniques.
[109]	The analysis of various machine learning algorithms for prediction of diabetes mellitus was given in the study with decision tree giving better results. No emphasis on missing value imputation and feature selection techniques was given.
[110]	In this study the author presents KNN, SVM and RF for diabetes prediction with RF giving most promising results. The downside of the study is lack of proper normalization techniques and study of limited algorithms.
[111]	The authors study application of SVM for early prediction of diabetes with emphasis of identification of individuals with high risk. The limitation of the study is dependence on single algorithm without comparing results of other algorithms.
[112]	The authors provide analysis of Naïve Bayes, Support vector machine, Random Forest and Simple Cart for early prediction of diabetes. The SVM shows better performance. However, the limitation of the study is consideration of limited number of techniques and evaluation metrics.
[113]	The authors gives comparative study of multiple machine learning algorithms and their contribution towards early diagnosis of diabetes mellitus. The limitation of the study is absence of missing value imputation that introduces bias in the results.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[114]	The authors propose a machine learning model to diagnose diabetes mellitus and other genetic disorders. The study mostly focuses on Random forest without considering other well known and advanced machine learning algorithms. Therefore, the study lacks in providing accurate picture of contribution of different algorithms in early detection of disease.
[115]	A novel approach for early prediction of hemorrhage using machine learning models was presented in the study that gives better accuracy, however, the drawback of the study is lack of validation of model and training of model on smaller size dataset.
[116]	A thorough comparison of various machine learning and regression models was given in study for the diagnosis of diabetes mellitus and shows the importance of using Electronic Health Records for training of models. The study shows edge of regression models over machine learning models in terms of interpretability and stability. The limitation of the study is the use of dataset without data preprocessing and data augmentation techniques.
[117]	The authors presents analysis of machine learning algorithms for early diagnosis of diabetes among which random forest gives better results in terms of accuracy. The downside of the study is the consideration of only few machine learning algorithms and with fewer evaluation metrics. Future research needs to address these limitation to get better results.
[118]	The study shows that ensemble approach of machine learning algorithms shows better performance in diagnosis of diabetes mellitus. The limitation of the study is that it does not take into account deep learning models.
[119]	In Type 1 diabetic patients the micro vascular complications are serious problems that need to be address. The study focuses the importance of early prediction of Micro vascular problems and the possibility of using machine learning in achieving this. The limitation of the study is lack of standardized measures and operationalization metrics.

TABLE 2.3: Limitations of various ML and DL based Methods

Ref	Limitations
[120]	A broad review of IoT based technology for diagnosing type 2 diabetes mellitus remotely was presented with the potential of wearable sensors to collect personalized data. The biggest limitation of the study is breach in data privacy and data security when the personalized data is shared publically by the wearable devices.
[121]	The study shows how random forest is more effective approach in early diagnosis of diabetes mellitus. The limitation of the study is implementation of only fewer algorithms and using datasets without preprocessing techniques that may produce biased results.
[122]	The study presents a novel approach for early prediction of diabetes mellitus using Convolutional LSTM. The downside of the study is application of Convolutional Neural Network on PIMA dataset which has small number of records and is not suitable for deep neural networks.
[1]	In this research, the authors present comparative study of basic machine learning algorithms for early diagnosis of diabetes. While SVM shows superior performance, the study limitation is that the dataset was used in its primitive form without application of data preprocessing techniques and feature engineering techniques. Besides the study does consider application of deep neural networks for early prediction of diabetes. Future research could focus on preprocessing, feature engineering, feature augmentation and use of deep neural networks for prediction of diabetes mellitus in order to improve the classification accuracy.
[123]	In this study the authors gives comprehensive review of various machine learning and fuzzy logic techniques for the early diagnosis of diabetes mellitus. From the study, it was found that fuzzy inference systems and random forest gives better prediction results. The limitation of the study is dependence on limited number of algorithms which does not give complete picture of the unsupervised ML algorithms for early prediction of diabetes mellitus.

The findings indicate that the Random Forests provide better accuracy followed by Gradient Boost and Logistic Reasoning. For early detection of diabetes mellitus, these algorithms may be considered while making efforts to keep in the queue the above-discussed insights for better performance and results. Deep Neural based algorithms

show promising results but fail to achieve higher accuracy without bias due to the smaller dataset size. A valuable insight into the problem is to use k-fold cross-validation techniques in tandem with a DNN to achieve promising results.

2.7.1 Experiment and Results

Most of the above-discussed techniques have used the PIMA dataset for their studies. A total of 9 attributes, that are significant to the process of prediction are used. With a total of 768 samples, 500 records are values containing 268 entries with a value of 1 (diabetic) and result attribute 0 (non-diabetic). Based on the results obtained in various studies done by researchers and discussed in the section, some of the most significant models have been implemented on the PIMA. The results are compared on various evaluation parameters to give thorough explanations of the optimality of these methods for diabetes prediction. The following ML techniques were implemented in Keras and TensorFlow:

1. KNN
2. SVC
3. LR
4. DT
5. GNB
6. RF
7. GB

The following evaluation criteria have been used to ascertain the performance of different ML methods: The accuracy and performance of any supervised learning method are examined using the confusion matrix (CM). Calculating the algorithm's performance using the following metrics:

$$Acc = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$Sen = \frac{(TP)}{(TP+FN)}$$

$$Spec = \frac{(TN)}{(TN+FP)}$$

$$F1\ Score = \frac{(2TP)}{(2TP + FP + FN)}$$

The CM is given in Figure 2.4:

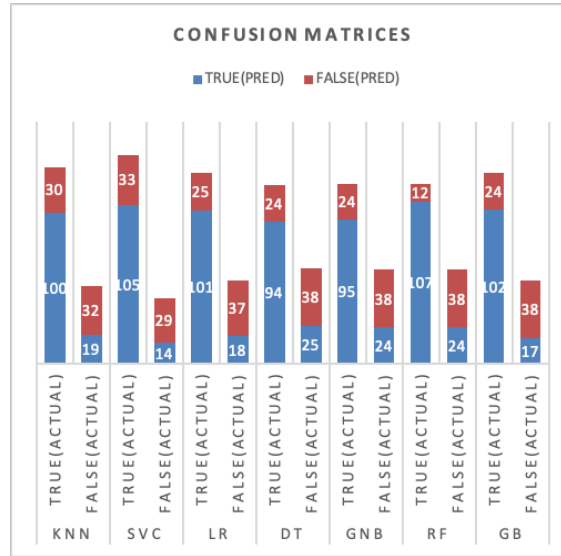


FIGURE 2.4: CM of various ML techniques

The analysis of ML-based methods reveals that some of the techniques provide better results than others. Based on this thorough comparative analysis, it was observed that few techniques including KNN, SVC, LR, DT, GNB, RF and GB provide results that are consistent among all studies. Subsequently, these techniques provide a promising roadmap to approaching this predictive problem. With this vision, the techniques were implemented for comparative analysis, as shown in Figure 2.4. The CM provides a clear summary of good classifiers for DM diagnosis. Consequently, the performance and rates of these ML-based techniques are given in Table 2.4 and Table 2.5 respectively. The RF algorithm performs fairly well in terms of performance and rates as well.

2.8 Conclusion

Many researchers use the PIMA dataset without applying pre-processing techniques such as normalization. Consequently, their findings are affected by outliers, overfitting, underfitting, and other anomalies. Additionally, some studies employ a limited range of

TABLE 2.4: Performance of various ML algorithms

Algorithm	Precision	Recall	Specificity	F1-Score	Accuracy
KNN	0.8403	0.7692	0.6275	0.8032	0.7293
SVC	0.8824	0.7609	0.6744	0.8171	0.7403
LR	0.8487	0.8016	0.6727	0.8245	0.7624
DT	0.7899	0.7966	0.6032	0.7932	0.7293
GNB	0.7983	0.7983	0.6129	0.7983	0.7348
RF	0.8992	0.8168	0.7600	0.8560	0.8011
GB	0.8571	0.8095	0.6909	0.8327	0.7735

TABLE 2.5: Rates of various ML algorithms

Algorithm	TPR	FNR	TNR	FPR
KNN	0.7692	0.2308	0.6275	0.3725
SVC	0.7609	0.2391	0.6744	0.3256
LR	0.8016	0.1984	0.6727	0.3273
DT	0.7966	0.2034	0.6032	0.3968
GNB	0.7983	0.2017	0.6129	0.3871
RF	0.8168	0.1832	0.7600	0.2400
GB	0.8095	0.1905	0.6909	0.3091

machine-learning algorithms for diagnosing diabetes and do not address missing values. There are also instances where the full potential of feature extraction is not utilized. Certain studies overlook the importance of all dataset attributes, particularly those like body size, height, and BMI, which significantly assist in the diagnosis of DM. This oversight negatively impacts the performance of classifiers. Several researchers had extensively studied diabetes mellitus, a life-threatening disease, due to its widespread grip on the world population. Many trials had been conducted to improve ML techniques for better accuracy. This study focused on analyzing and comparing various techniques to uncover their limitations and drawbacks. Many parameters, such as missing values, inadequate datasets, inefficient feature extraction, reduced biomarkers, and medication effects on significant parameters, were often disregarded while using ML and DL classifiers for diabetes mellitus diagnosis.

The results of various studies indicate that Random Forests provide the highest accuracy, followed by Gradient Boosting and LR. For the early diagnosis of DM, these algorithms should be considered, along with the incorporation of the aforementioned insights, to achieve better performance and results. There is a great potential of these AI-based techniques in aiding diabetes diagnosis, and their ability to analyze large datasets and uncover hidden characteristics that traditional approaches might overlook. A comparative analysis of ML and DL techniques, outlining the strengths and limitations in the context of diabetes diagnosis indicated how various ML methods, including DT, SVM,

ANN, and K-NN perform and the potential of DL techniques such as deep belief networks and RNN were explored. Based on the background and the survey done and the promising results while working on DNN based approaches, it was inevitable to focus on developing a model for DM diagnosis using DL techniques while focusing on pre-processing of the dataset.

Chapter 3

Enhancing Diabetes Mellitus Diagnosis - A Comparative Analysis of Pre-processing Techniques for Data Optimization

3.1 Introduction

Pre-processing is necessary before using the dataset to train any ML or DL model. Pre-processing techniques can be used for imputation of missing values, standardization, normalization, feature reduction, etc. There are many techniques available for imputation of missing values, and the most popular ones are mean, mode, and median. These are straightforward imputation techniques, but they have some serious drawbacks, such as the fact that applying these techniques to the dataset always introduces some bias, which may lead to poor performance from ML and DL models. To address these issues with biased datasets, more complex and advanced techniques for data imputation—such as polynomial regression—have been used. Following the process of imputation of missing values, elimination of any extraneous characteristics from the dataset in order to cut down on the duration needed to train the model is done. This is accomplished by reducing the dataset's dimensionality through the scoring of attributes based on how much of a contribution each attribute makes to the classification process. A variety of techniques are available for this purpose, and Spearman's rank correlation coefficient has been used.

3.2 Conventional machine learning techniques

Diabetes is a kind of long-lasting disease and tremendous amount of data is generated during the treatment of patients suffering from diabetes like cardiovascular disease and the early prediction could prove lifesaving. Various ML methods can be utilized for the diagnosis and prediction of DM, however, these techniques can either be supervised or unsupervised forms of learning.

3.2.1 Supervised Learning

A goal function is set that defines the model and then fine-tune arguments of the classification technique by the application of a group of recognized sample categories. The list of identified output variable values is used to compose training data. Supervised learning has two types of learning tasks, classification, and regression. Some of the popular supervised learning methods include SVM, ANN, and DT [72].

3.2.1.1 Decision Tree (DT)

Based on true or false responses, the decision tree categorizes the data. The final structure is represented graphically as a tree, an acyclic graph made up of nodes and edges. There are three different kinds of nodes: the root node, which is a single node; internal nodes, which have one parent and one or more children; and leaf nodes, which are also known as external nodes and have no children. A decision tree has an advantage over other machine learning algorithms in that it is simple to comprehend. Selecting a feature from the dataset to act as the decision tree's root node is the first stage in the constructing process. The final class was typically not precisely predicted by a single trait. We call this impurity. To determine how exactly a given feature classifies the supplied data, as well as to compute the impurity level, a variety of functions such as Gini, entropy, and information gain [134] can be utilized. The node with the least amount of impurity is chosen at any level. When dealing with numerical characteristics, the average of neighbouring values is found and the data is sorted in ascending order to ascertain the Gini impurity. GI is computed at every selected average value by arranging data points depending on whether the value of features is lower or higher than the selected value and whether the selected value classifies the data correctly. The equation used to find Gini impurity is given as:

$$Gini\ Impurity = 1 - \sum_{i=1}^k p_i^2$$

where k =number of classification categories p =proportion of instances of those categories. For each value of leaf nodes, the weighted average of Gini impurities is calculated and the value with the smallest impurity is selected for that feature. The same process is repeated for all features to select the feature and value that will be selected for the node. For each node the process is iterated at every depth level until all the data is classified. On the creation of the tree, the prediction can be done by going down the tree using various conditions at each node to complete the classification Figure 3.1.

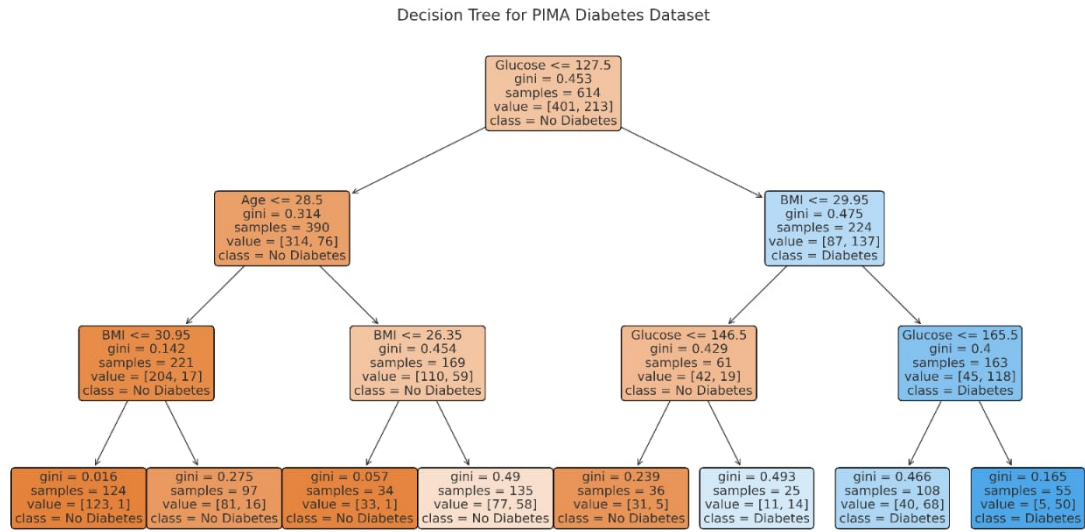


FIGURE 3.1: Decision Tree classification of PIMA dataset

3.2.1.2 Support Vector Machine (SVM)

A kernel function is used by SVM to transform data from input space to multidimensional feature space, after which it looks for a separating hyper-plane. PUK, radial basis functions, and linear polynomial kernel functions are some of the frequently utilized kernel functions in SVM. It is a popular supervised ML technique that is used for regression and classification [135]. Classification is the use of SVM that occurs most frequently. In multidimensional space, the SVM method determines the ideal hyperplane that separates data points into distinct classes according to their properties. To maximize the margin between two nearby data points, the hyperplane is used. The number of features in feature space determines the number of dimensions of the hyperplane. The hyperplane will be a 2-D plane when there are two features in feature space. Suppose we have

one dependent variable, represented by either a red or blue circle as shown in Figure 3.2, and two independent variables, x_1 and x_2 .

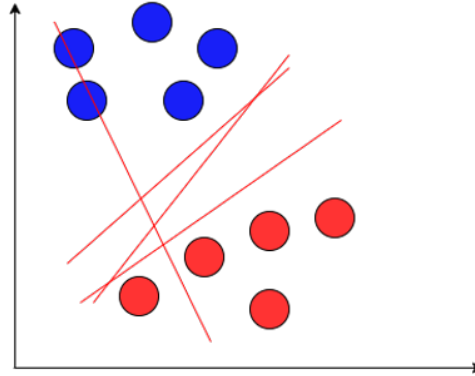


FIGURE 3.2: Linearly Separable Data Points

Figure 3.2 illustrates this with many lines representing our 2-D hyper-planes. These are caused by the two features x_1 and x_2 , which divide our data points into two sets of circles. As can be seen in the above picture, there are several lines that divide our data points into red and blue circles (the hyperplane in this case is a line since only two input features, x_1, x_2), as can be clearly seen. Thus, how can we select the optimal line, or more broadly, the optimal hyperplane, to divide our data points?

1. Working: The hyperplane that separates the data points with large margins is considered a good hyperplane. A hyperplane with a maximum distance between it and the nearest data point on both sides is the best option. It is referred to as the maximum-margin hyperplane/hard margin if one exists. The ideal hyperplane is L2, as seen in Figure 3.3.

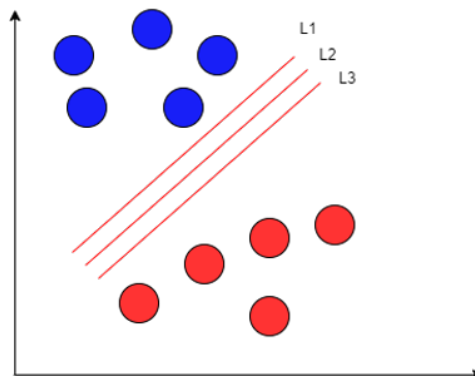


FIGURE 3.3: Multiple Hyper planes

Thus, the hyperplane with the greatest distance between it and the closest data point on each side is selected. It is referred to as the maximum-margin hyperplane/hard margin if one exists. Therefore, we select L2 from the given diagram. If the scenario is different, as shown in Figure 3.4:

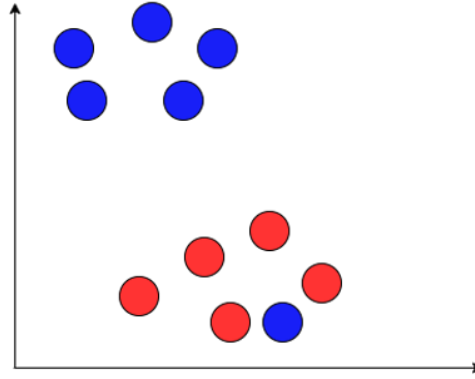


FIGURE 3.4: Selecting Hyper plane for data with outlier

In this case, one blue data point is in the boundary of red data points. In such cases, the blue data point in the red boundary is known as the outlier of the blue data point. In order to maximize the margin, the support vector machine ignores the outlier and chooses the optimal hyperplane. This makes the support vector machine resilient to anomalies. Similar to the previous example, the SVM determines the biggest margin and further applies a penalty each time a data point crosses the margin Figure 3.5. The margins in these situations are referred to as "soft margins." Should the dataset have a soft margin, the objective of SVM is to minimize $(1/\text{margin} + (\sum \text{penalty}))$.

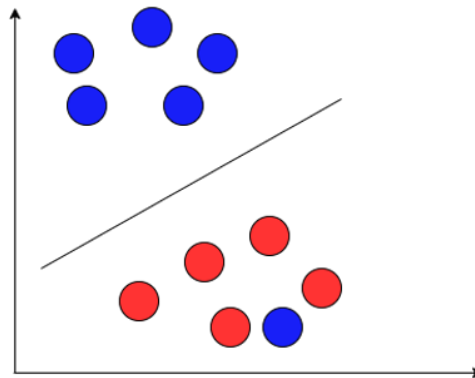


FIGURE 3.5: Hyperplane which is the most optimized one

In all the above cases the data points were linearly separable i.e., the data points were separable by a straight line. It is not always the case. There can be data which is not linearly separable as shown in Figure 3.6 below:

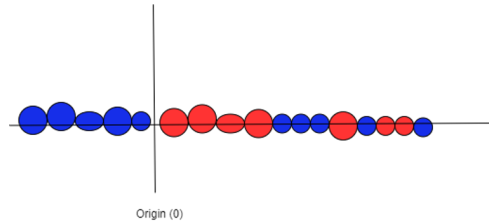


FIGURE 3.6: Original 1D dataset for Classification

In the above diagram, the data is not separable linearly. The SVM handles this problem by creating a variable known as kernel. A point x_i data point on the line and another variable that depends on how far away the origin o is taken. The plot is shown in Figure 3.7 below:

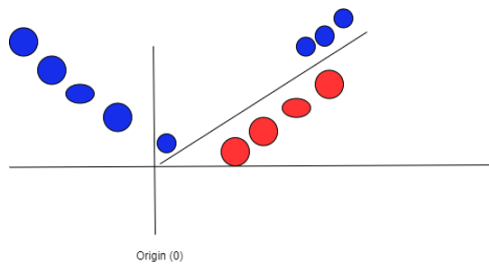


FIGURE 3.7: Mapping 1D Data to 2D

A non-linear function called a kernel is employed to use the new variable y as a function of distance from the origin.

3.2.1.3 Artificial Neural Network (ANN)

ANNs are among the most potent instruments available for analyzing complicated clinical information. ANNs predict dependent values from a given set of independent values after detecting the complicated link between dependent and independent variables using known data during the training phase. The ANN is suitable for incomplete, complex, and non-linear datasets [10].

3.2.1.4 K-nearest neighbor (KNN)

Thomas Cover created the KNN method for classification and regression. The K-closest training examples for input in the feature space make up this non-parametric method, and its output depends on whether it is employed for classification or Regression [136]. This technique produces class membership as output which is an object classified on plurality vote. The object that is the most common among the neighbors is allotted to the class. It is simply assigned to the class of the one nearest neighbor object if $k=1$. The property value of the object, which is the mean of the values of the k nearest neighbors, is the result of the k -NN regression. This supervised machine learning approach is widely used and may be applied to both regression and classification tasks. Evelyn Fix and Joseph Hodges created the method in 1951, and Thomas Cover later made modifications to it. This algorithm is most extensively used in pattern recognition, data mining and intrusion detection problems. In KNN there is some predefined data known as training data that helps in classification of coordinates into groups recognized by an attribute. Let us consider the following data points with two features Figure 3.8.

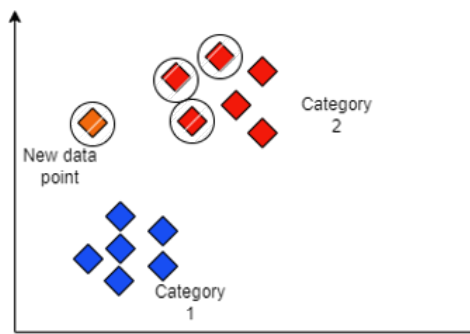


FIGURE 3.8: KNN algorithm working visualization

There are two categories of data points known as Category 1 (blue) and Category 2 (red). We also have a new data point called testing data and the objective is to assign this new data point to one of the two categories. Initially, all the new data points which are not classified are marked as white. If all the points are plotted on a graph, it is possible to identify some clusters or groups. On receiving new data point which is unclassified, we can allocate it to a group by observing its characteristics and deciding what group its nearest neighbors belongs to. A new data point close to the group of points classified as red (category 2) has more probability of being classified into category 2.

3.2.2 Unsupervised Learning

Enormous data is produced by the medical industry, but some data doesn't possess corresponding labels. Therefore, it is inappropriate to employ supervised learning approaches. In such cases, unsupervised learning can be useful to discover the relation between hidden structures of unlabeled data. Some of the commonly used unsupervised learning methods include clustering techniques and association rule techniques [137].

3.2.2.1 Clustering Techniques

The clustering techniques can be used for searching for useful patterns in unorganized datasets.

3.2.2.2 Association Rule Learning

An association rule often has the form $X_1, X_2, \dots, X_n \rightarrow Y$. An association rule is one where the two entities support degree and trust degree reach a threshold value that is already specified. Commonly used association rule discovery algorithm includes Apriori.

3.3 Deep Learning Techniques

The traditional ML algorithms have limitations in handling raw natural data. To deal with such type of data, the researchers use the notion of feature learning or representation learning in which there is a set of techniques that are used by a machine and provided with input in the form of raw data and it learns the representations required for the classification [10]. One of the widely used representations learning methods is Deep-Learning that has many layers of representation, obtained by a combination of simple nonlinear modules that helps to transform a simple level of representation into a higher, more abstract representation. The datasets generated by the medical industry are multidimensional, heterogeneous, and sparse. Therefore, the utilization of DL methods for the classification of such datasets is more suitable and has applications in the prediction of diabetes [138].

3.3.1 Convolutional Neural Network

One of the Deep Neural Networks in Artificial Intelligence that is most frequently researched and utilized is the Convolutional Network. CNN has several uses in computer vision, natural language processing, and image processing. Among the popular feed-forward neural networks for image processing CNN is the leading Deep Neural Network. It is composed of several Multi-Layer Perceptrons. Hubel and Wiesel researched the visual cortex of cats and found that the visual cortex of cats has a kind of hierarchical information processing structure in which the information processing complexity increases with an increase in several levels Figure 3.9.

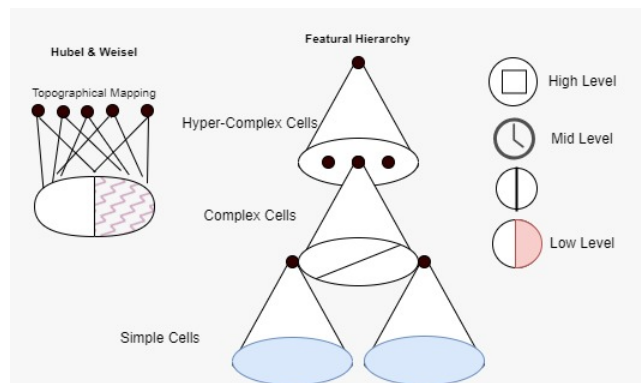


FIGURE 3.9: The Hierarchical Cortical Model of Cat's Visual Cortex

Yann and LeCun inspired by the Hubel and Wiesel observations proposed the CNN model. The first working CNN model commonly known as LeNet was proposed by Yann and LeCun with a back propagation algorithm and random gradient descent method. This model marks the foundation for CNN which was then popularly used for image processing and image recognition applications. The CNN has several similarities with biological neural networks. Each Convolutional Layer in the CNN model has receptive cells known as neurons that are connected with the input data. They work similarly to biological neurons. Each CNN model has five layers. One or more convolutional layers that are utilized for calculations come after the input layer, which is the first layer. Activation functions are found in the third layer; the pooling layer comes next; and the output layer, commonly referred to as the fully linked layer, is the last layer. The input data or images are read by the data input layer in which the pre-processing of data is performed like normalization, standardization, and feature selection. This first layer facilitates further image processing operations.

The Convolutional Layer is the most important layer that performs feature extraction. This layer performs two main operations viz local cross-validation and receptive field.

The size of the receptive field is determined by the convolutional kernel size and the network configuration. In different levels, the receptive field's dimensions also differ. The size of the receptive field also decides how much information in input data can be observed by the neurons which ultimately affects the learning ability and performance of the network. Once the filter size, stride and fill values are fixed, the convolutional layers perform the calculations to extract various features and reduce the noise impact in the convolutional computation layer. The third layer contains activation functions to map the results of the output layer to the convolutional layer. The most commonly used activation functions are Rectified Linear Unit and Softmax. In the fourth layer known as the pooling layer, the dimensionality reduction is done using feature selection to reduce overfitting without changing features. The two-dimensional feature map is transformed by a fully connected layer into a one-dimensional vector with the application of parameter optimization and changing weights.

Grid-like topological data, such as photographs, is processed using Convolutional Neural Networks, or ConvNets. A two-dimensional binary representation grid structure is used to represent an image, with each cell holding the pixel value that indicates the color and brightness of each pixel in the image as shown in Figure 3.10.

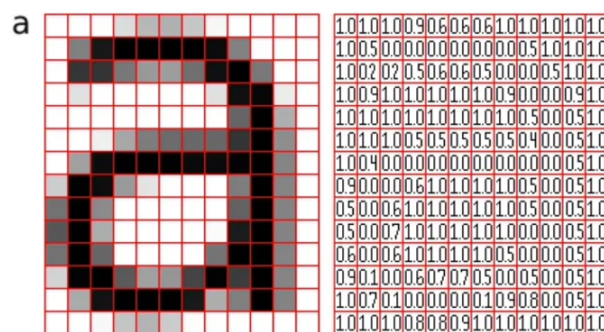


FIGURE 3.10: Representation of image as a grid of pixels

The CNN has fundamental units that perform computation on the input data known as neurons. Only the receptive field—a specific region of the image—is processed by each neuron in a CNN. CNN layers are organized such that more complex patterns, like as faces, letters, and objects, are recognized in advanced levels, while simpler patterns, such as lines and curves, are discovered early.

3.3.2 CNN Architecture

The basic CNN has three layers: Convolutional, pooling and a fully connected layer Figure 3.11.

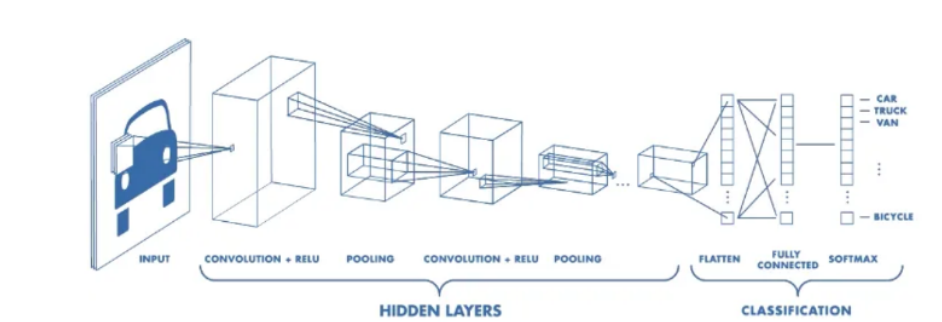


FIGURE 3.11: Architecture of CNN

3.3.2.1 The Convolutional Layer

The convolutional layer is the fundamental component of the CNN model. Most of the computations made by the CNN model are handled by this layer. Two matrices are used as input in this layer. The first matrix is made up of a set of learnable parameters called the kernel, and the second matrix is made up of the pixel values of the receptive field—a specific area of the input picture. In terms of depth, the kernel is larger than the picture, but its size is lower overall. Put another way, if our image has three RGB channels, the kernel's height and width will be tiny, but its depth will encompass all three channels Figure 3.12.

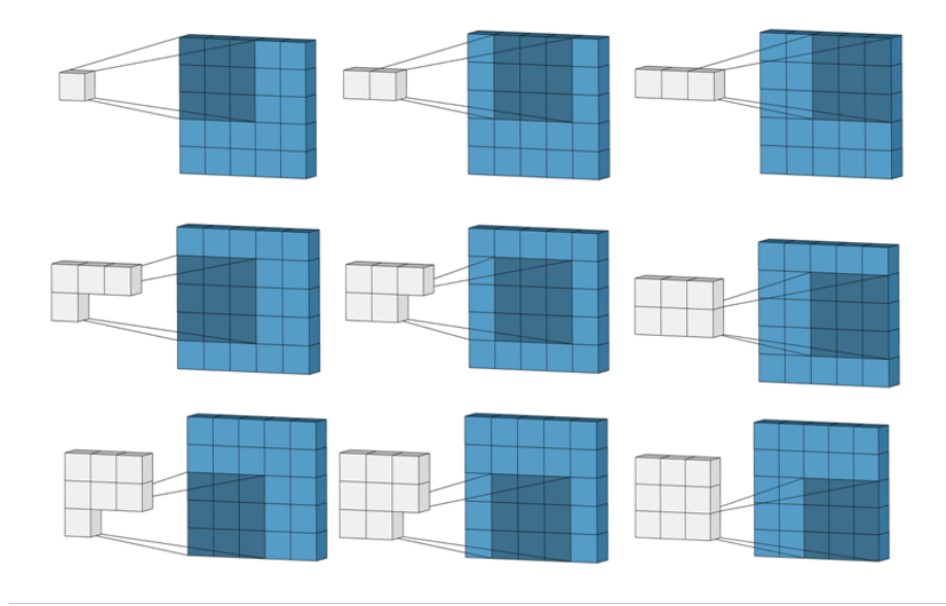


FIGURE 3.12: Illustration of Convolution Operation

During the forward pass, the kernel moves across the picture's height and breadth to produce an image representation of the specific receptive region of the input image. This forward pass results in the creation of an activation map, a two-dimensional representation of the picture. The number of cells that the kernel slides is known as the stride. Let us assume an image of $W \times W \times D$ dimensions and D_{out} kernels. Let f be the spatial size, S be the stride and P the amount of padding, the formulae for calculating the size of output volume is:

$$W_{out} = \frac{W - F + 2P}{S} + 1$$

The output volume of size $W_{out} \times W_{out} \times D_{out}$ will be produced using the aforementioned formulas. Three benefits come from the convolution process: equivariant representation, parameter sharing, and sparse interaction. A typical neural network applies matrix multiplication of parameter matrix and finds the relation between input and output unit. Each input unit interacts with each output unit. In contrast, a Convolution Neural Network uses sparse interaction by making the kernel smaller as compared to the input size. For example an image consists of millions of pixels but for the purpose of processing only tens or hundreds of pixels can contain relevant information, which is detected by the CNN Kernel. The number of parameters that need to be stored is minimized which not only reduces memory requirements but also improves the learning rate and other statistical efficiency of the model. In conventional neural network designs, every weight

matrix value is utilized just once and is never reused. However, in the context of CNN, shared parameters are utilized, meaning that the weights applied to one area of the network are the same as the weights given to other sections in order to create output. As a result of parameter sharing, the CNN layers have the equivariance to translation property i.e., on changing input, the output also gets changes similarly Figure 3.13.

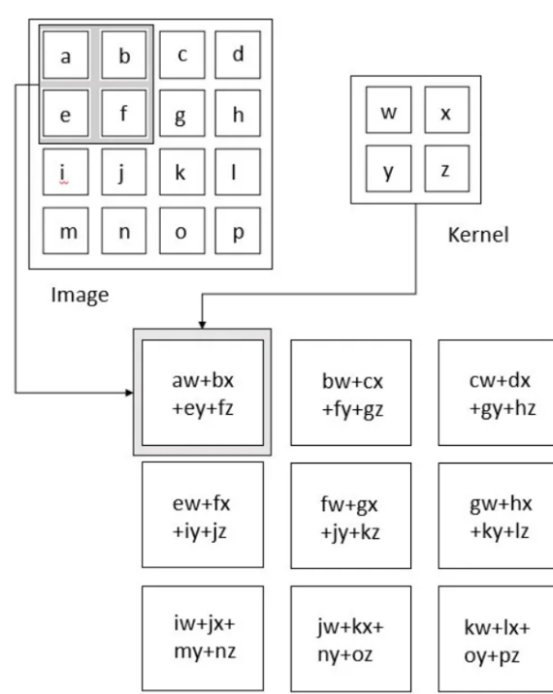


FIGURE 3.13: Convolution Operation

3.3.2.2 Pooling Layer

The pooling layer's primary goal is to minimize the number of computations and weights needed. This is accomplished by substituting different statistical summaries, such as an aggregate of the closest outputs, for the network's output at particular points. The representation's spatial size is therefore decreased. Every slice is subjected to the pooling process independently. A weighted average depending on the distance from the center pixel can be employed, as can an average of the neighborhood rectangular area, the neighborhood rectangular area's L2 norm, or both. The most often used pooling function is max pooling, which provides the neighborhood's maximum output Figure 3.14.

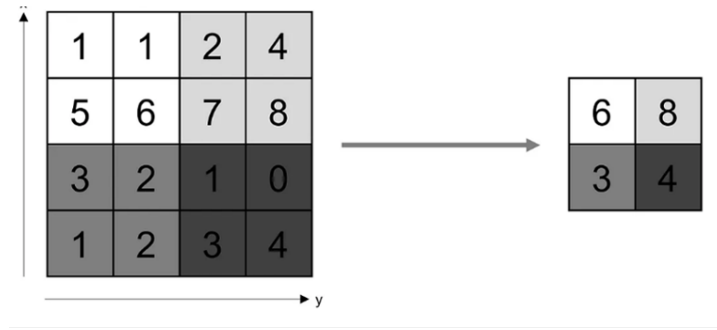


FIGURE 3.14: Pooling Operation

A pooling kernel with spatial sizes F and S as the Stride size and an activation map of size $W \times W \times D$ will produce an output volume of W_{out} . These values are determined via formula:

$$W_{out} = \frac{W - F}{S} + 1$$

3.3.2.3 Fully Connected Layer

In a completely linked layer, the matrix multiplication process is followed by the bias effect, and every neuron in the previous layer is connected to every neuron in the subsequent layer. Determining the representation between the input and the output is the primary goal of the fully linked layer.

3.3.2.4 Non-Linearity Layers

Non-linearity layers are positioned immediately following the convolutional layers in order to add non-linearity to the activation map. There are a number of non-linear operations that can be applied. The mathematical formula for the sigmoid function is:

$$\sigma(k) = \frac{1}{(1 + \varepsilon(-k))}$$

A real-valued number is transformed into a range between 0 and 1 via the sigmoid function. The gradient of the sigmoid function decreases to almost zero at each tail of activation, and if the gradient is too tiny, this is the negative portion of the function, the gradient may be missing during the backpropagation. Another drawback of the sigmoid function is that if the input data of the neuron is always greater than zero, the sigmoid

output will be either positive always or negative always. This results in a weight gradient update that is zigzag in dynamic.

Real-valued values are converted using the Tanh function to fall between -1 and 1. Tanh's output is always centered at zero.

ReLU is the most popular non-linear function used over the past few years. The ReLU has mathematical formula $f(k) = \max(0, k)$. Compared to other nonlinear functions, ReLU's activation is more dependable since it has a zero threshold. The ReLU converges six times faster than Sigmoid and Tanh. The downside of ReLU is that it can be fragile when the model is being trained. If a gradient with large gradient is passed through it, the update is done in a manner that the neuron never gets updated. This problem can be eliminated by setting a proper learning rate.

3.3.2.5 Models of Convolutional Neural Network

There are several models that are utilized that are based on convolutional neural networks for image processing and the most popular are LeNet, AlexNet, GoogleNet and ResNet. Among all Convolutional Neural Networks, LeNet is the earliest CNN model that was originally developed to recognize handwritten digits. The LeNet is a five-layer structure with simple architecture. The LeNet has a small Kernel and uses a Sigmoid function in the activation layer that produces binary output. Therefore the LeNet provides relatively simple processing as compared to other CNN models Figure 3.15.

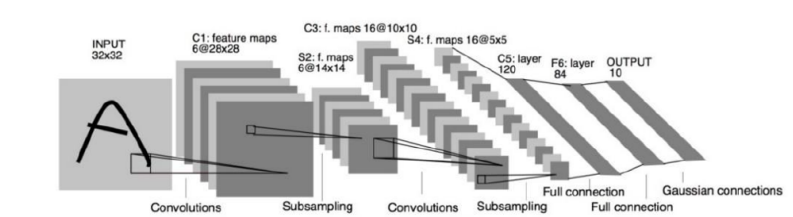


FIGURE 3.15: Working of LeNet

Alex Krizhevsky along with his associates developed a model in 2012 namely AlexNet which is considered to be the first model that uses deep neural networks for image processing applications. AlexNet is a more complex network as compared to LeNet. The AlexNet features three fully connected layers in addition to five convolutional layers. Its excellent feature extraction capabilities allow it to learn more. It is more generalized deep neural network with better performance. It uses a more advanced ReLU activation function in the activation layer and also has one extra normalization layer and dropout technique that helps in the reduction of overfitting. In the last two convolutional layers,

the AlexNet has convolutional kernels of varied sizes for feature extraction and training with images of different sizes Figure 3.16.

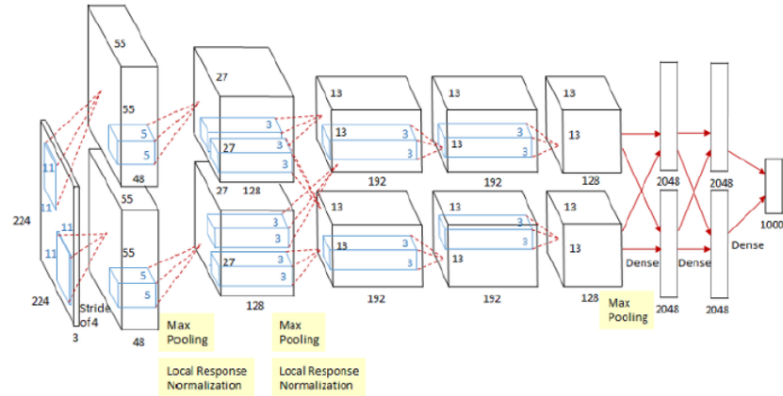


FIGURE 3.16: Network Structure of AlexNet Model

In 2014, the Google Brain team proposed an advanced Convolutional Neural Network known as GoogleNet which used an inception module for feature extraction of various scales by the application of parallel convolution kernels of various sizes and pooling operations. In GoogleNet the memory consumption and computation required were reduced to a large extent by the application of small convolution kernels and by using global average pooling that reduces the count of attributes in the network. Consequently, the performance of GoogleNet increases dramatically. The introduction of batch normalization technology in GoogleNet improves its stability and helps to accelerate the convergence process of the model. It also decreases model dependency on initial weights Figure 3.17.

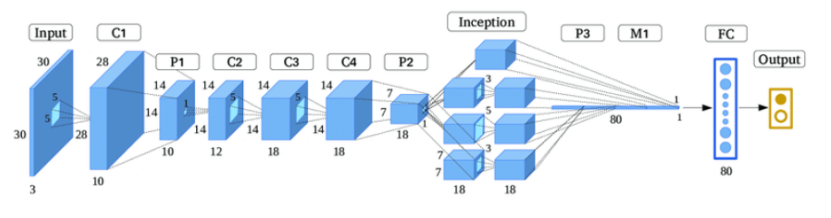


FIGURE 3.17: Network Structure of GoogleNet Model

3.3.2.6 Residual Network (ResNet)

After the first CNN-based Network popularly known as AlexNet was developed, there was a notion that any successful deep neural network architecture can lower the error

rate by increasing the number of layers. This was true as far as the number of layers were smaller, but when we start adding more layers to a deep neural network the problem known as Vanishing or Exploding gradient is introduced that results in either zero gradient or gradient becomes overly large as a result of which the training as well as testing error rates goes on increasing with the increase in number of layers Figure 3.18.

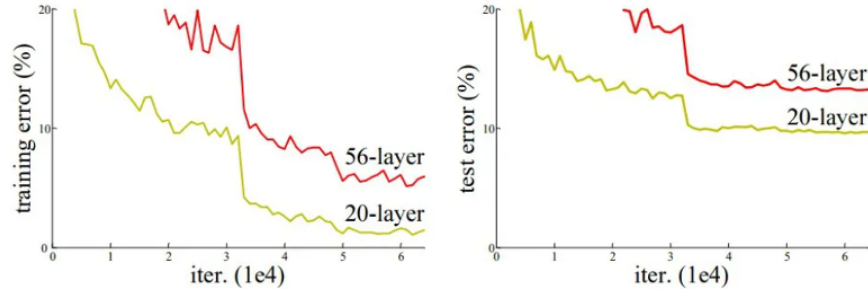


FIGURE 3.18: Comparison of 26 layer VS 56 layer architecture

It is clear that 26-layer CNN architecture performs better as compared to 56-layer CNN architecture on training and testing data. The researchers from various experiments and observations concluded that this increase in error rate is caused by vanishing or exploding gradient due to deep neural networks with more layers.

To address this vanishing or exploding gradient problem, Microsoft Research experts introduced a novel architecture in 2015 known as residual network or ResNet. This network solves the vanishing gradient problem by employing a method called skip connections. A leftover block is created in the skip connection Figure 3.19 by bypassing parts of the levels that occur between link layer activations to succeeding layers. A stack of such leftover blocks forms resnets. The resnet's concept is to let the network fit the residual mapping so that layers may pick up the underlying mapping. The general formula for skip connection is:

$$F(x) = H(x) - x, \text{ where } H(x) = F(x) + x$$

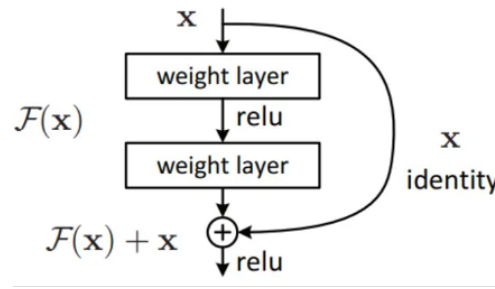


FIGURE 3.19: Skip Connection

The benefit of a skip link is that any layer which deteriorate the network performance will be skipped by regularization and hence we can train extremely deep neural networks without the problem of vanishing gradient.

3.3.2.7 ResNet Architecture

A 34-layer plain architecture used by ResNet with shortcut/skip connections was developed and named VGG-19. VGG-19 architecture is converted into the residual network with the help of these shortcut connections as shown in Figure 3.20.

3.3.3 Deep Belief Network

DBN is another classification mechanism that specifically consists of Restricted Boltzmann Machines with hidden layers interconnected with a visible layer of the next RBM. Deep Belief Network has also produced promising results in the classification of diseases and can be advantageous in the detection of diabetes [126]. One of the classy ANN derived from Machine Learning algorithms is DBN. Through the automatic discovery of patterns from big datasets, this type of deep learning network is utilized to extract significant information. It is a multi layered network and each layer can extract information from the data obtained from previous layer and builds a complex understanding of overall data in the dataset. The multiple layers of the Deep Belief Network consist of various layers of stochastic units known as restricted Boltzmann machines. The objective of various layers of DBN is to extract various features hidden in input data. The lower layers are responsible for identifying basic or simple patterns while as higher layers are to recognize abstract concepts Figure 3.21. In this way, DBN is effective in learning complex representations of data in the dataset. DBNs are suitable to be used with unsupervised learning where the data has no predefined labels.

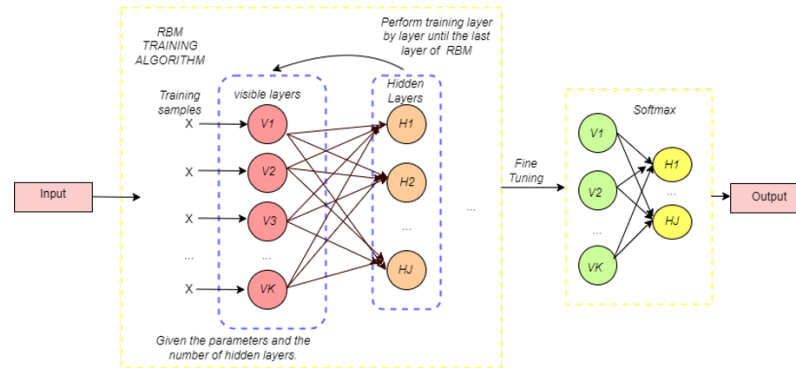


FIGURE 3.21: Deep Belief Network Structure

There are two main phases for DBN to accomplish its task. Pre-training is the first stage, and fine-tuning is the second. In the pre-training stage, the input data is understood by the network layer by layer. Each layer works as an independent RBM and learns complex data more effectively. During the first phase, how likely it is that the input data will occur is understood by the network layers to get insight into the structure of the input data. During the fine-tuning stage, the network's parameters are fine-tuned or adjusted for particular tasks like regression or classification. Through the use of the network's backpropagation mechanism, these parameters are adjusted. The performance is evaluated and the parameters adjustment is done based on errors in the network.

3.3.4 Recurrent Neural Network

RNNs, which allow the output from the previous layer to enter the next phase, are among of the most popular types of neural networks [139]. The inputs from this layer and the output from the layer before it are not reliant on one another in other neural networks. The RNN maintains the state with the help of hidden layers Figure 3.22. The most important concept of RNN is a hidden state whose aim is to maintain the information about the sequence. The memory state is another term for this maintenance stage. The memory state keeps track of the prior input that was sent to the network. Every input has the identical parameters, and all inputs or hidden layers undergo the same processing to yield the same outputs. This reduces the complexity of the parameters.

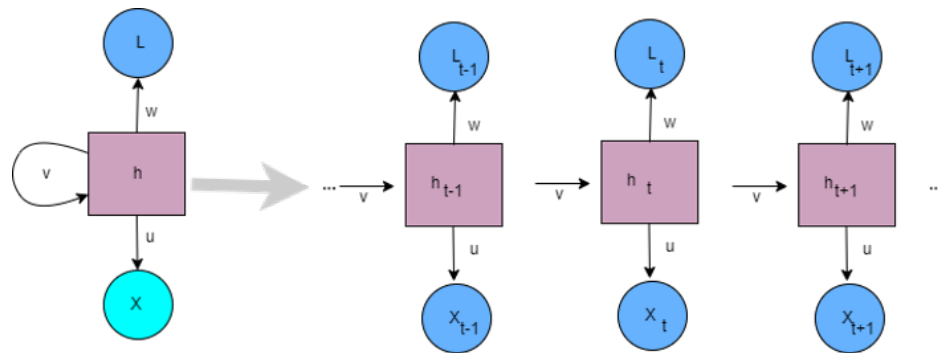


FIGURE 3.22: Recurrent Neural Network

Artificial Neural Networks without any looping nodes are known as feed-forward networks. Also known as a multilayer neural network, feed-forward networks transfer data from the input layer to the output layer via hidden layers without creating loops. With no feedback loops, it is not possible for the feed-forward network to store or retain the previous inputs and is therefore less significant with sequential data analysis. Other classifications of DL are given in Figure 3.23.

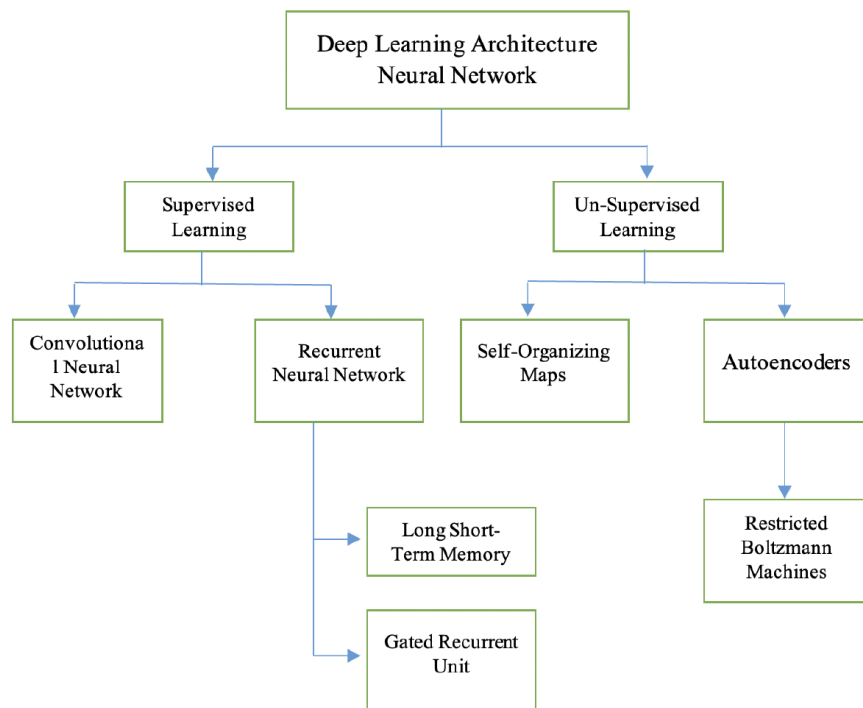


FIGURE 3.23: Deep Learning Classification

3.4 Experiment and Results without preprocessing

The most popular methods for data mining were applied. The PIMA dataset, which comes from the National Institute of Diabetes and Digestive and Kidney Diseases, was utilized for this work. It was used in its primitive form without application of any preprocessing techniques. Table 3.1 shows the Attributes Chosen for Study and the findings are shown in Table 2.4 and Table 2.5.

TABLE 3.1: Attributes

S No	Attribute Name
1	Diastolic Blood Pressure
2	Plasma Glucose Concentration
3	Number of Times Pregnant
4	Body Masss Index
5	2-Hr Serum Insulin
6	Triceps Skin Fold Thickness
7	Age
8	Diabetes Pedigree Function
9	Class(Yes or No)

The results suggest that the Random Forests provide better accuracy followed by Gradient Boost and Logistic Reasoning. For early detection of diabetes mellitus, these algorithms may be considered while making efforts to keep in the queue the above-discussed insights for better performance and results. But because the experiment was conducted using a crudely constructed dataset, it includes a large number of missing values, unnormalized, imbalanced, and superfluous data. Using the dataset without handling these anomalies may give biased results. To eliminate such problems, data preprocessing may be considered to transform the original dataset into the suitable form for better accuracy and unbiased results.

3.5 Dataset

One well-known dataset that is often utilized in statistics and machine learning for carrying out research on diabetes diagnosis is the Pima dataset. It includes information gathered from the Pima Indian community close to Phoenix, Arizona, about female Pima Indians [140]. It includes a range of medical measurements as well as information on each person's development of diabetes within five years of the measurements. Figure 3.24 shows the description of the dataset.

```
dataframe.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

FIGURE 3.24: Description of the PIMA Dataset

Below is a summary of its characteristics:

1. Pregnancies: Total count of pregnancies.
2. Glucose Concentration: Plasma glucose levels measured during a two-hour OGTT.
3. Blood Pressure: Measured in millimeters of mercury, the diastolic blood pressure.
4. Skin thickness: The Triceps skin fold thickness (mm).
5. Insulin: Serum insulin (μ U/ml).
6. BMI: Weight in kg/(height in m)².
7. Diabetes pedigree function: A method that uses family history to determine a person's risk of having diabetes.
8. Age: Age of patient in years.
9. Result: 1 if the individual has diabetes, 0 otherwise.

3.5.1 Advantages

1. Relevance to the real world: The dataset is suitable for healthcare analytics and forecasts since it is based on actual medical data [59].
2. Widely used: Because of its popularity, there is an abundance of available documentation, tutorials, and research, which makes it simpler for beginners to comprehend and utilize [141].

3. Diversity of features: It has a range of features that allow for the investigation of many elements that may be connected to diabetes, perhaps leading to the development of more thorough models [141].

3.5.2 Disadvantages

1. Small sample size: By today's machine learning standards, the sample size is small, with about 700 instances. Over-fitting may result from this, particularly in complicated models [50].
2. Absent information: A few entries have missing values; these could need to be imputed or handled differently during the analytical process [142].
3. Potential bias: Models developed using the dataset might not have good generalization to other populations because it is specific to the demographic of Pima Indians. When this is applied to different populations or ethnic groups, it may produce biased results [51].

3.5.3 Age of Dataset

The dataset was collected in the 1980s, so it may not fully represent the current medical landscape or include newer factors relevant to diabetes diagnosis or prediction [143]. The Figure 3.24 gives some insights for each attribute. It is clear that the dataset suffers from number of anomalies like missing values and un-normalized values [125]. Before the dataset can be utilized, it is necessary to pre-process it so that the anomalies will be eliminated [55]. The distribution of various attributes of the dataset can be shown using histogram and density plots in Figure 3.25 and Figure 3.26 as:

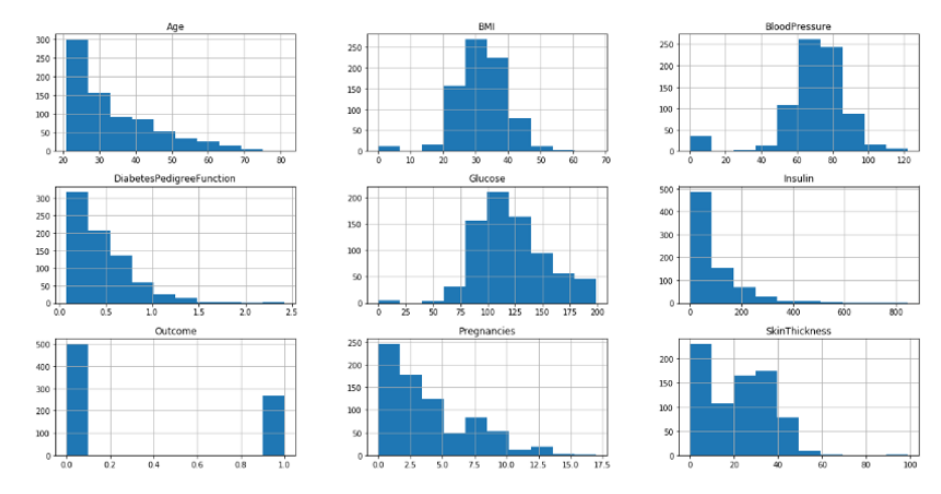


FIGURE 3.25: Attribute distribution of PIMA Dataset using Histogram

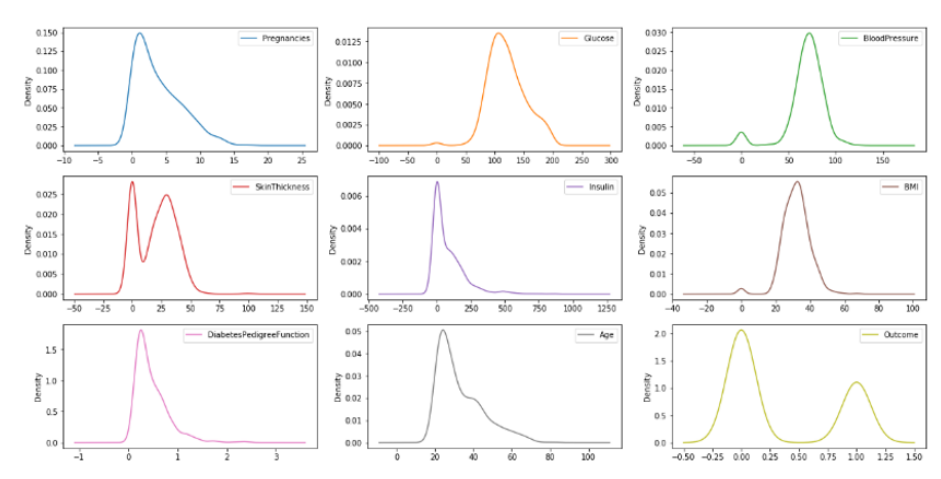


FIGURE 3.26: Attribute distribution of PIMA Dataset using Density plots

3.6 Preprocessing

Preparing raw data to make it appropriate for model training is known as preprocessing in data science. Since the model's performance and accuracy are heavily impacted by the caliber of the input data, this is a crucial stage in ML and DL [144]. The pre-processing includes series of operations to make the data clean and error free and modify the raw data into the form that is appropriate for the ML model's training. The various preprocessing methods that are frequently used in machine learning are given below:

3.6.1 Data Cleansing

3.6.1.1 Managing Missing Values

An experiment or observation is a statistical technique for optimizing the performance of a system with some input variables. This experiment starts with an investigational design for a trial plan with all known factors affecting the system's result. The input data collected in a well-planned experiment and under a completely controlled manner may still contain some missing data that can affect how well the system functions significantly and reduce the statistical power of the system and the system may produce biased results or sometimes inaccurate results. Dealing with the missing values is the most important challenge during an experiment. Erasing the records entirely is the simplest way to handle missing values. However, if a substantial portion of the dataset is missing, this might lead to the loss of other important data. Imputation is another widely used technique for handling missing value problems. In this way, the system may generate more accurate and efficient outputs by substituting values for the missing values. If the dataset has missing value it may greatly impact the performance of model in negative way. Missing values can be addressed by deleting the rows or columns that contain the missing data, imputation i.e. replacing missing values with computed estimate, and use sophisticated methods to find missing values like interpolation are some of the strategies [59]. Various missing value techniques used in this study are:

1. Mean: The mean of the available values in that column should be used to complete any values that are not present. It is suitable for data that is continuous. It may cause the distribution to be distorted in the event of outliers. In the dataset, for each column C, the mean x'_j of all values available x_{ij} in column C is calculated. Then the missing value x_{ij} is replaced with the mean value x'_j [145].

$$x'_{ij} = \begin{cases} x'_{ij} & \text{if } x_{ij} \text{ is missing} \\ x_{ij} & \text{otherwise} \end{cases}$$

2. Mode: Replace any missing values with the column's mode, or the value that appears the most frequently. Adequate for discrete or category data. Unsuitable for numerical data that is continuous. In the dataset, for each column C, the mode_j (most frequent value) of all values available x_{ij} in column C is calculated.

Then the missing value x_{ij} is replaced with the value $mode_j$ [146].

$$x'_{ij} = \begin{cases} mode_j & \text{if } x_{ij} \text{ is missing} \\ x_{ij} & \text{otherwise} \end{cases}$$

3. Median: The column's median, or the middle value after the data is sorted, should be used to complete any values that are not present. Suitable for data that is continuous. Robust to outliers compared to mean imputation. In the dataset, for each column C, the median x'_j (middle value) of all value available x_{ij} in column C is calculated. The missing value x_{ij} is replaced with the median x'_j [147].

$$x'_{ij} = \begin{cases} x'_j & \text{if } x_{ij} \text{ is missing} \\ x_{ij} & \text{otherwise} \end{cases}$$

4. Polynomial Regression: It is important to note that when basic approaches like mean, median, and mode are used to manage missing values, the dataset may include bias of some kind which leads to inaccurate prediction. In order to handle these limitations an intriguing method known as polynomial regression can be used for missing value imputation [127]. In order to use polynomial regression for missing value imputation following steps are performed:

- (a) Data Preparation: Divide your dataset into two sections: one including all of the data, and the other containing values that are missing and will require imputation.
- (b) Model Training: Consider the missing value feature to be the dependent variable and the other characteristics to be independent variables for each feature that has missing values. Utilizing all of the data, train a polynomial regression model.
- (c) Prediction: To predict the missing values for each feature, use the trained polynomial regression model.
- (d) Imputation: Use the predicted values from the polynomial regression model to complete missing values.
- (e) Evaluation: Analyze the model's effectiveness using a validation set or compare the imputed values to the true values, if available, to determine how well the imputation performed.

While the independent variable(s) in PR employ concepts that are polynomial, the formula itself is structured similarly to that of linear regression [148]. The

following formula can be used to illustrate a basic case of univariate polynomial regression, which is polynomial regression with a single independent variable [127]:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n + \varepsilon$$

With x acting as the independent variable and y acting as the dependent variable. $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the coefficients, where the coefficients for the polynomial terms of x are $\beta_1, \beta_2, \dots, \beta_n$, and the intercept is β_0 . The error term, denoted by ε , is what encapsulates the variation between the values that are seen and those that are anticipated [49].

3.6.1.2 Handling Outliers

Outliers are data points that significantly deviate from the rest of the dataset. They have the power to distort model behavior and statistical metrics. To control their influence, strategies such as winsorization, cutting, or outlier transformation can be used [149] [46].

3.6.2 Data Transformation

3.6.2.1 Normalization

The normalization scales numerical features to a predefined range usually between 0 and 1 or -1 and 1 so that every feature contributes equally to train the model. The two most used normalizing techniques are Min-Max and Z-score scaling.

1. Z-score: It is one of the popular preprocessing methods used to rescale numerical data in order for the data's SD to be one and its mean to be zero. Each data point's mean is deducted, and the remaining amount is divided by the attribute's standard deviation. The formula below may be used to get the Z-score of a data point x in a feature with mean μ and standard deviation σ . [4]:

$$z = \frac{x - \mu}{\sigma}$$

Where: x = original value of the data point, μ = mean of the feature, σ = standard deviation of the feature and z = standardized value, or Z-score, of the data point. The z-score ensures that every feature in the dataset have a distribution having mean of 0 and standard deviation of 1. The z-score transformation ensures that all

features are on the same common scale. The mean of each feature will be centered around 0 after standardization and the standard deviation will be 1. This helps to make the feature comparison and interpretation easier [150]. The downside of Z-Score is that it may increase the effect of outliers on the scaled data and thus might not be suitable for features with non-Gaussian distributions or in situations where outliers are present [58].

2. **Min-Max Scaling:** Min-max is a normalization technique used to rescale numerical values to some fixed range, usually between 0 and 1. Using Min-Max, the value for each feature is mapped proportionately. The minimum value gets mapped to 0 and the maximum value to 1 with all other values between 0 and 1. The min-max scaling formula for a data point x in a feature with a minimum value of $\min(x_i)$ and a maximum value of $\max(x_i)$ is [28]:

$$x_{scaled} = \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)}$$

Where: x = original value of the data point, $\min(x_i)$ = minimum value of the feature, $\max(x_i)$ = maximum value of the feature, x_{scaled} = scaled value of the data point. Each feature in the dataset will have values that fall between $[0, 1]$ when min-max scaling is applied to it. Neural networks, gradient descent-based optimization techniques, and methods that rely on distance measures (e.g., k-nearest neighbors) can all benefit from this normalization. These algorithms all require the input features to be on a similar scale [47]. While scaling all values within a predetermined range, min-max scaling maintains the relative distances between data points and the distribution's form. It may, however, struggle to deal with outliers because it condenses the range of most data points into a narrower region, which could result in information loss. Furthermore, skewed or non-Gaussian characteristics may not be a good fit for min-max scaling. Min-max scaling is a simple and useful technique for normalizing data to a specified range, which makes it easier to read and compare across characteristics, despite certain restrictions [57].

3.6.2.2 Encoding Categorical Variables

Categorical variables must be converted into a numerical representation since machine learning models usually requires numerical input. Binary, label, and one-hot encoding are examples of common approaches.

3.6.3 Feature Selection

Machine learning algorithms are based on the axiom "Garbage in, garbage out," where "garbage" is defined as unwanted information or data. For better ML, a sizable dataset is used to train the model. There is a lot of noise in this massive dataset, and some of the attributes in the dataset aren't useful for categorization. Large amounts of data can stifle the learning capacity of the model, and including unnecessary data can lead to an erroneous prediction. Feature selection is a mechanism that shrinks the input vector of our model by eliminating noisy data from the dataset and trains the model only on significant attributes. It automatically removes the irrelevant features from the dataset that do not contribute towards the classification process so that the resulting model is more accurate and learns in less time. The concept of feature selection is given in Figure 3.27.

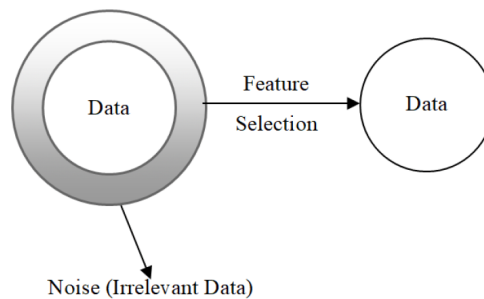


FIGURE 3.27: Feature Selection

3.6.3.1 Feature Selection Models

The Feature Selection models are supervised and unsupervised Figure 3.28. The supervised feature selection uses labels for the output. In this method, the target variable is used to find the attributes that can enhance the performance. In the case of unsupervised feature selection models, there is no need for the output label class. The feature selection is done based on unlabelled data.

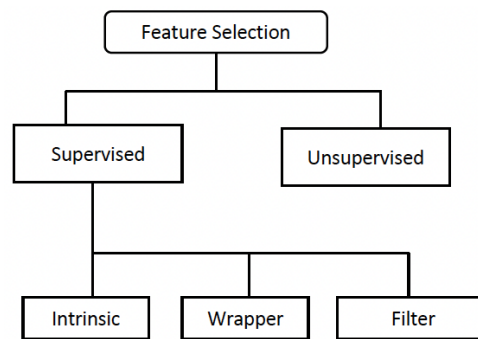


FIGURE 3.28: Feature Selection Methods

1. **Filter Method:** This method of feature selection works on the concept of correlation as shown in Figure 3.29. Coefficient of determination between dependent and independent characteristics is computed using the technique. It checks which features are positively related to the output variables and which features are negatively related to the output variable. Based on this the positively related features are considered and negatively related features are dropped from the dataset. Information gain and chi-square are among the filter-based feature selection techniques.

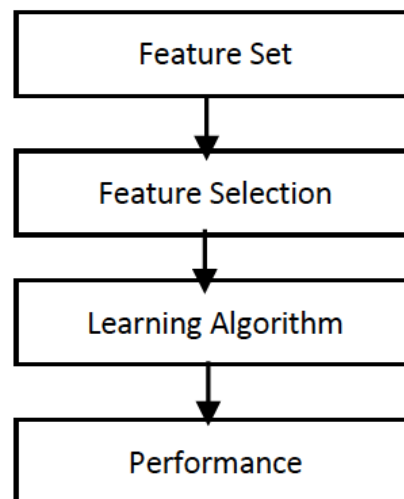


FIGURE 3.29: Filter Feature Selection

2. **Wrapper Method:** In this feature selection method Figure 3.30, a portion of the data is extracted, and the model is trained using this portion of the data. Considering the learning model's output, the features are added or subtracted and the training process is restarted. This method of feature selection, known as greedy feature selection, assesses the efficiency of every conceivable feature combination.

Examples of Wrapper methods of feature selection are Forward Selection and Backward Elimination.

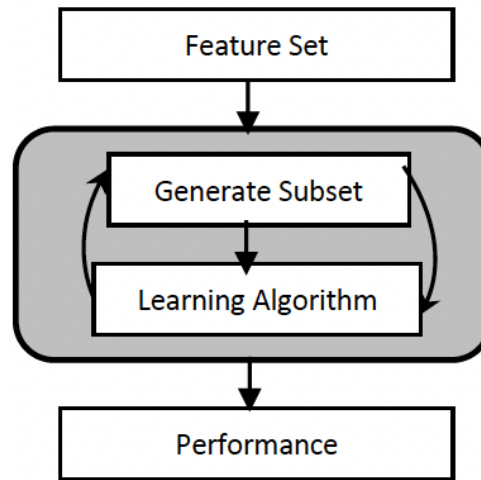


FIGURE 3.30: Wrapper Feature Selection

3. **Intrinsic Method:** The Intrinsic method is the combination of Filter and Wrapper techniques and utilizes both mechanisms to create the best subset of data for the learning algorithm Figure 3.31. The Intrinsic method trains the model iteratively while trying to maintain the computation cost at a minimum. Examples of intrinsic methods are Lasso and Ridge Regression.

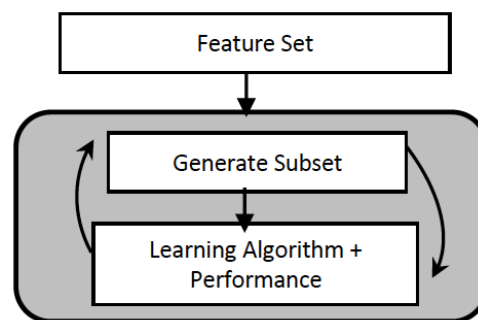


FIGURE 3.31: Intrinsic Feature Selection

3.6.4 Feature Engineering

3.6.4.1 Developing Derived Features

To simplify intricate relationships among the data or to capture more pertinent information, new features can be created from pre-existing ones. This may entail domain-specific

information, interaction terms, or mathematical transformations [151].

3.6.4.2 Dimensionality Reduction

In high-dimensional datasets, the issues of over-fitting and higher computational costs are inevitable. Principal Component Analysis (PCA) and feature selection approaches are examples of dimensionality reduction techniques that can be used to minimize the amount of features while maintaining critical information [152]. The choice of FS or DR depends upon the type of attributes whether the attributes are of categorical nature or numerical type. Table 3.2 gives the description of various techniques that can be applied for dimensionality reduction [64].

TABLE 3.2: Various feature selection techniques

In Parameter	Out Parameter	Suitable Model for feature selection
N	N	Pearson's Spearman's
N	C	ANOVA (linear) Kendall's (nonlinear)
C	N	Kendall's (linear) ANOVA (nonlinear)
C	C	Chi-Squared test Mutual Information

As far as PIMA dataset is concerned, all the attributes are numerical in nature, therefore the most suitable choice is Pearson's Coefficient or Spearman's Coefficient [18]. The decision between two depends upon whether the relation between dependent and non dependent attributes is monotonic or not. Most of the independent attributes of PIMA dataset are monotonically related with the dependent attribute(outcome). Therefore, the most suitable model to be used to feature selection is Spearman's rank correlation coefficient [24]. A metric used to quantify the degree and direction of correlation between two ranked variables is the Spearman's rank correlation coefficient. It's frequently applied when there may not be a linear relationship between the variables under study or when the data is ordinal rather than interval. SRC coefficient can be helpful in determining correlations for feature selection in machine learning or statistics [69]. This is one possible usage for it:

1. Ranking: Prioritize your features and your target variable independently. Spearman's rank correlation coefficient functions on ranked data, which makes this significant [53].

2. Calculation: The SRC coefficient between each feature and the target variable should be calculated. This shall provide an estimate of how closely each attribute is related to the objective [52].
3. Selection: Features that exhibit strong monotonic correlations with the target variable are indicated by high absolute values of the Spearman's rank correlation coefficient. one should consider include these characteristics in the model [56].
4. Validation: As with any feature selection strategy, it is important to make sure that the features that have been chosen actually improve the performance of the model and the features are not merely artifacts of the specific dataset that is being dealt with [137]. To do this, one could use techniques like cross-validation [60].

The formula for SRC coefficient (ρ) is as follows:

$$\rho = \frac{6 \sum d^2}{n(n^2 - 1)}$$

where d is the distinction in the relevant variables' rankings and the number of observations is denoted by n [10].

3.6.5 Normalization of Data

3.6.5.1 Handling Skewed Data

Model performance can be negatively impacted by skewed distributions, especially for methods that are sensitive to the data's distribution. To improve the symmetry of data distributions, methods like the Box-Cox transformation or log transformation can be applied.

3.6.6 Data division

3.6.6.1 Train-Test Split

To assess model performance, the dataset is divided into separate testing and training sets [24]. This guarantees an efficient evaluation of the model's potential to generalize to unknown inputs.

3.6.7 Managing Imbalanced Data

Imbalanced datasets with a significantly higher number of one class than the other are classes that might provide biased models in classification tasks [30] [28]. To solve this problem, methods such as oversampling, undersampling, or the application of algorithms (like SMOTE) developed to deal with imbalanced data might be used.

Preprocessing, in general, is essential to machine learning because it guarantees that the input data is accurate, pertinent, and properly organized for efficient model training and assessment [32] [29]. Depending on the characteristics of the dataset and the requirements of the used learning technique, each preprocessing step should be carefully selected.

3.7 Experiment and Results with Preprocessing

To check the significance of preprocessing techniques on model performance, number of machine learning algorithms was executed including artificial neural network. Table 3.3 shows the accuracy of various classifiers without using any preprocessing techniques. Table 3.4 and Table 3.5 summarizes the results obtained from pre-processing techniques in Artificial Neural Network.

TABLE 3.3: Accuracy (Different Classifiers) without pre-processing Techniques

Model	Accuracy		
	Mean	Median	Most Frequent
Naïve Bayes	75.58	69.05	75.57
Random Forest	77.36	75.57	75.41
KNN	72.31	73.61	72.96
SVM	77.04	76.21	77.04
Decision Tree	70.36	67.43	75.57
ANN	71.66	58.50	62.89

The various algorithms were used in which trivial missing value imputation techniques viz mean, median and mode, were used for replacing missing values in dataset. As these techniques has the tendency to introduce bias in the dataset and ultimately result in overfitting. During the experiment no standardization or normalization techniques were used. The effect of application of standardization/normalization techniques on artificial neural network is shown in Table 3.5. In this experiment, Z-Score and Min-Max techniques were used for scaling along with mean, mode and median as missing value imputation techniques.

TABLE 3.4: Artificial Neural Network performance after pre-processing Techniques

Missing value strategy	Z-score	Minmax scaler
Mean	75.75%	84.77%
Median	60.89%	82.14%
Most frequent	65.19%	82.79%

The table 7 below summarizes the results of the Neural network when

1. Used with primitive dataset without application of any pre-processing technique
2. Used with mean, median and mode as missing value imputation techniques
3. Used with mean mode and median as missing value imputation techniques along with Z-Score standardization technique
4. Used with mean mode and median as missing value imputation techniques along with Min-Max standardization technique

TABLE 3.5: Application of missing value imputation and normalization on Artificial Neural Network

Model	Dataset	PreProcessing	Technique used	Accuracy
ANN	PIMA dataset in its primitive form	No Preprocessing	None	76.9%
ANN	PIMA dataset with missing value imputation	MVI	Mean	71.66%
			Median	58.50%
			Most Frequent	62.89%
ANN with Normalization	PIMA dataset with normalization	Z-Score	Mean	75.75%
			Median	60.89%
			Most Frequent	65.19%
		Min-Max	Mean	84.77%
			Median	82.14%
			Most Frequent	82.79%

The experiment shows the effect of pre-processing techniques on medical datasets, such as Diabetes Mellitus, using various machine learning techniques is significant in improving accuracy. Various machine learning models were compared using different missing value strategies in the dataset, and the ANN was used to predict Diabetes Mellitus in the missing values dataset using the pre-processing techniques that include z-score and MinMax. The results showed that the ANN accuracy is significantly improved using the pre-processing techniques.

3.8 Classification using DNN

Deep neural networks (DNNs) are artificial neural networks (ANNs) that include many hidden layers positioned between the input and output layers. A DNN's hidden layers

process the input data through a number of changes to extract progressively abstract features. These networks function especially well for tasks like audio and picture identification, and more due to their capacity for learning intricate patterns and representations from input. When training deep neural networks, methods such as optimization algorithms and backpropagation are usually used to reduce the difference between the predicted and actual outputs, adjust the weights and biases of the network.

3.8.1 Methodology for proposed DNN Model

Figure 3.32 describes the methodology of a DNN model for Diabetes Diagnosis:

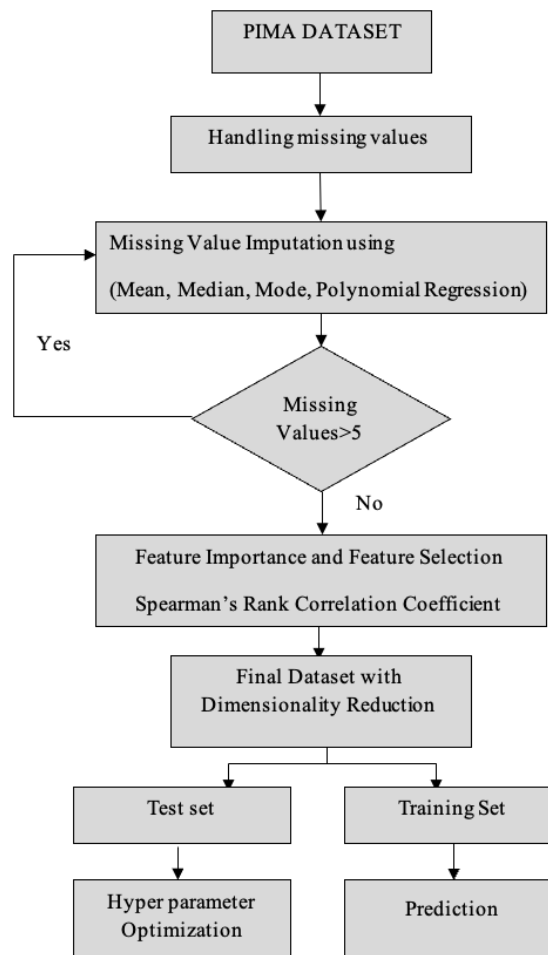


FIGURE 3.32: Proposed Methodology for a DNN based Model

3.8.2 Experiment and Results

The experiment was done on the PIMA dataset. The flowchart in Figure 3.33 was followed in this experiment.

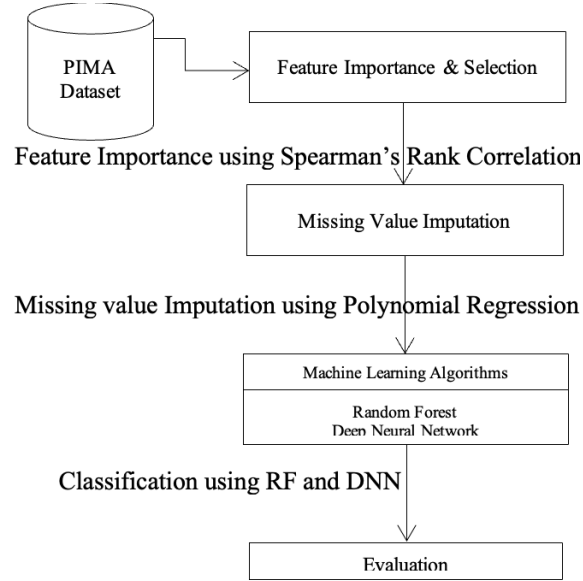


FIGURE 3.33: Proposed Methodology

In the first part of the experiment, Random Forest (RF) and the proposed DNN model were trained with no feature selection (NFS). Then the same models were trained with selected features using Spearman's rank coefficient to obtain and compare the outcomes as given in Table 3.6.

TABLE 3.6: Performance (%) with and without Feature Selection

Metric/Classifier	RF	DNN	RF	DNN
	NFS	NFS	WFS	WFS
Precision	88.65	96.21	96.65	97.35
Recall	92.50	96.00	96.50	96.67
F1 Score	90.43	96.05	95.98	96.97
Train Acc	92.00	100.00	97.38	98.71
Test Acc	92.50	96.00	96.50	96.67

As is evident from Table 3.6, DNN based model performs better than RF for both the cases NFS and WFS. It is also clear that feature selection models have higher performance than NFS models. The performance findings are given in Figure 3.34 below:

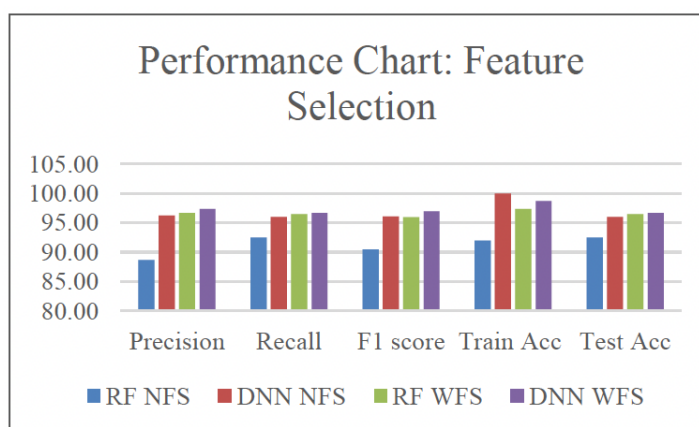


FIGURE 3.34: Feature Selection Performance (%)

The suggested DNN model was trained using feature selection and missing value imputation techniques in the second portion of the experiment, and the results were compared and displayed in Table 3.7.

TABLE 3.7: Performance (%) with Mean, Median and Polynomial Regression

Metric/Technique	Mean	Median	Polynomial Regression
Precision	97.05	97.05	98.12
Recall	96.75	96.75	97.93
F1 Score	96.76	96.76	97.95
Train Acc	98.21	98.21	98.62
Test Acc	96.75	96.75	97.93

The performance findings are given in Figure 3.35 below:

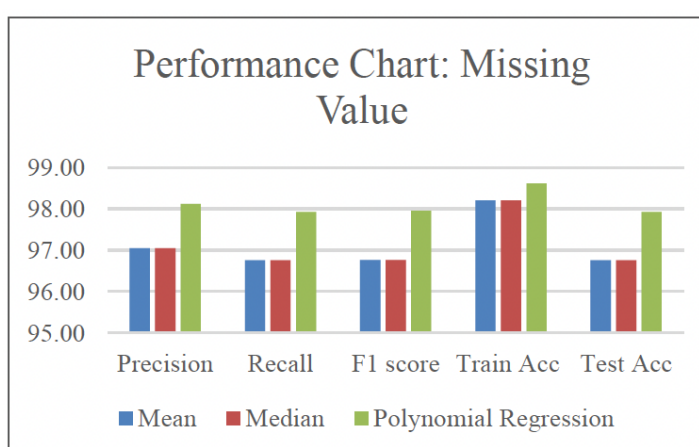


FIGURE 3.35: Missing Value Performance (%)

DNN-based model WFS performs better with polynomial Regression, as summarized in Fig. 4 than any of the other missing value imputation techniques.

3.9 Conclusion

Contributing to the increasing amount of research on DM diagnosis, the effectiveness of data pre-processing techniques in enhancing Deep Neural Network (DNN) performance is critical. A DNN-based model was proposed that integrates feature selection at various stages of the prediction process, resulting in a robust and reliable framework. This model outperformed a Random Forest model, achieving an accuracy of 96.00% without feature selection and 96.67% with feature selection. The impact of missing value imputation indicated that the DNN model achieves an accuracy of 97.93% when employing polynomial regression for this task. These findings highlight the significance of FS and MVI in improving the accuracy of DNN models for early DM diagnosis.

However, the research also identifies limitations associated with DNNs trained on text datasets with limited data records. The study suggests that DNNs might not be fully utilized in such scenarios, perhaps resulting in less than ideal performance. To tackle this, leveraging large datasets, particularly image datasets, for DNN training may be proposed. This approach could be further enhanced by data augmentation techniques like flipping, rotation, transformation, and scaling. By artificially adding more examples to the quantity of training samples, these techniques can potentially enhance the efficiency and robustness of DNNs in diagnosing DM. This aligns with the broader field of medical diagnosis using machine learning, where pre-processing techniques like z-score and MinMax normalization have been demonstrated to greatly increase the accuracy of models like Artificial Neural Networks (ANNs) for predicting DM, especially in datasets with missing values. Overall, this study emphasizes the significance of methods for pre-processing data for DNNs in diagnosing DM and emphasizes the potential areas for future research to further optimize model performance, especially the need for an image dataset for the application of a DL-based model.

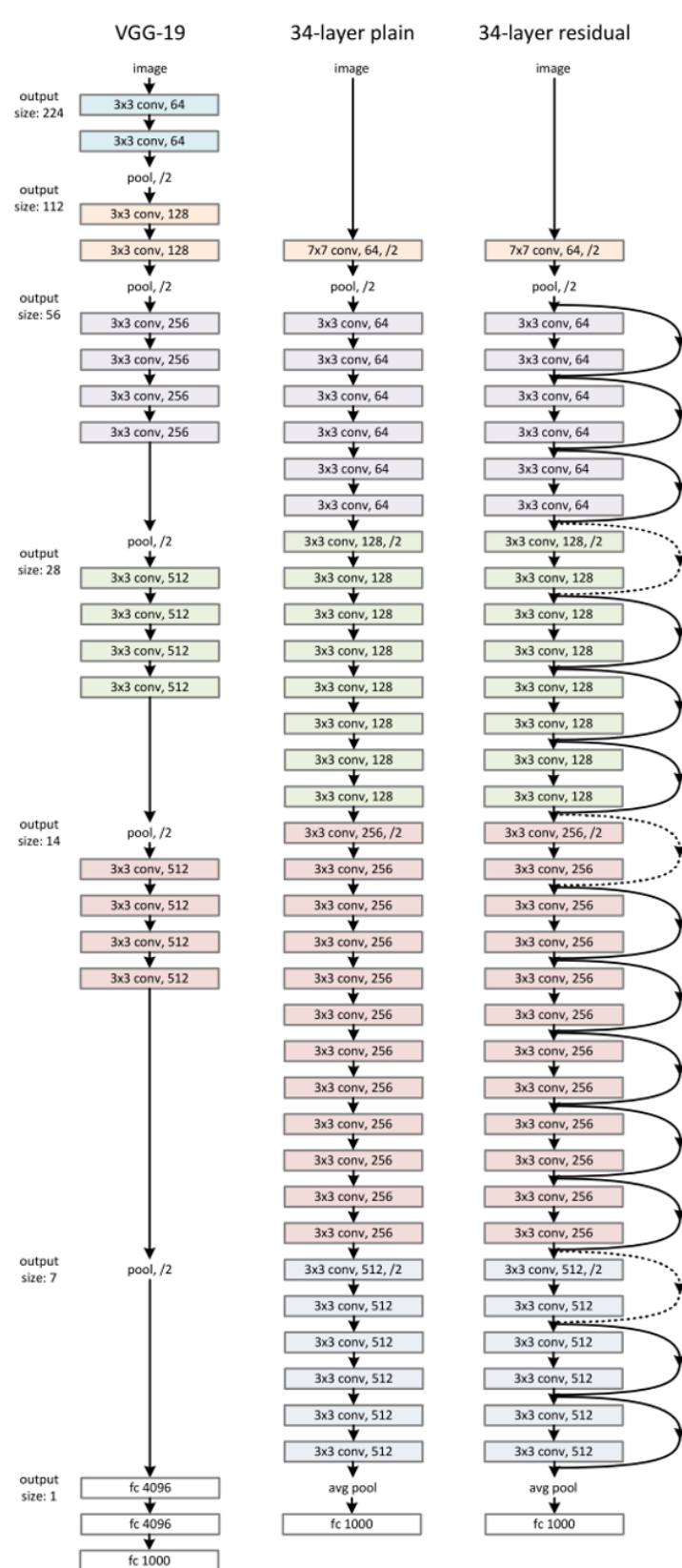


FIGURE 3.20: ResNet 34 Architecture

Chapter 4

Novel Approach to convert text-based PIMA dataset into image dataset

4.1 Introduction

Deep learning is a highly useful technique for the early identification of diabetes mellitus, according to the study done by numerous authors over the past few decades. By using pre-processing techniques on the dataset to get rid of various anomalies like over-fitting, under-fitting, redundancy, missing values, and non-significant features to make it more efficient for analysis, it is possible to increase the effectiveness of deep learning algorithms for diagnosing the disease. The work addresses the global problem of diabetes by exploring a revolutionary deep-learning method for early identification. Conventional convolutional neural network (CNN) models have drawbacks when used with numerical medical datasets, like this study's PIMA Indians Diabetes Database. A technique for transforming numerical data into visual representations depending on feature relevance is needed to get over this obstacle. This conversion makes it possible to use strong CNN models for early diabetes diagnosis.

4.2 Proposed Methodology

Figure 4.1 describes the proposed methodology for early diagnosis of DM using Image dataset:

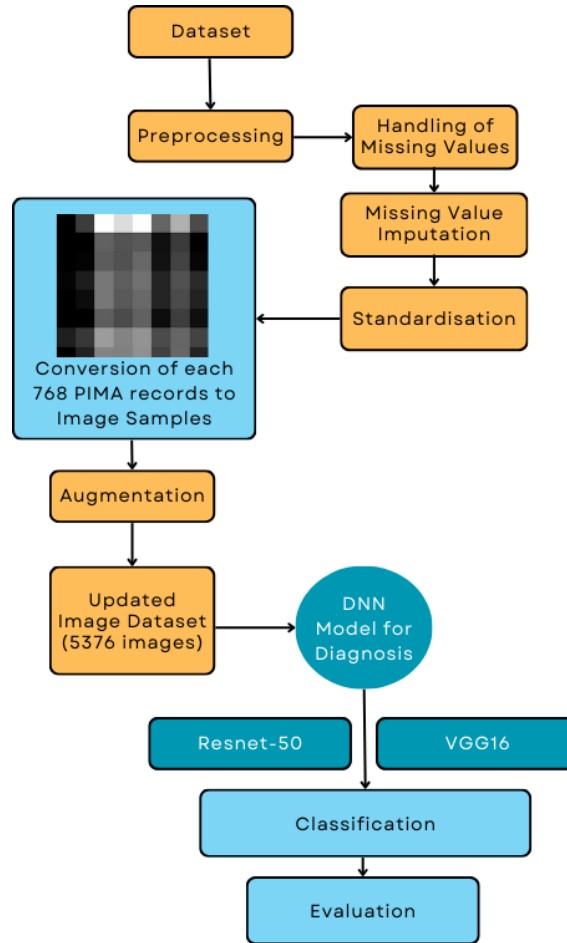


FIGURE 4.1: Propose Methodology

4.2.1 Handling of Missing Values

The result of an experiment depends upon the input dataset. When the data is collected in an organized manner, the model trained on that dataset produces reliable results [153]. However, most of the data is collected in controlled environment that may cause biased results. Due to the presence of anomalies like missing or non-available values, un-normalized data, redundant attributes the model may give biased classification and hence is unreliable. In order to deal with such anomalies in the dataset, number of techniques has been developed. For example, to address the problem of dataset's missing

values, the simplest approach is to delete the records having missing values in any of the attributes. However, if the number of missing values in the dataset is large, then this technique of removing non-available records from the dataset may lead to information loss. Another technique to deal with missing values is to use basic data imputation techniques like mean, mode and median. When replacing missing values with the help of these statistical observations, there is always scope for introducing bias in the dataset that may ultimately lead to biased results. Therefore, some advanced missing value imputation techniques can be applied like KNN, Regression and Hot-deck.

1. Mean Imputation: In this method, the missing value of any feature in a particular record is replaced by the mean value of other records for that attribute. In this way, the sample size of the dataset is preserved and it is easy to use, however, it reduces the variability in data which may lead to underestimation of standard deviation and variance. The mean imputation is calculated as follows:

$$\hat{z}_i = \bar{z}_h$$

Where \hat{z}_i = imputed value of record i and

z_h = sample mean of respondent data within some class

2. Regression: In order to deal with the problems faced in mean imputation method, regression imputation can be used. The formula for missing value imputation using regression is as follows:

$$\hat{z}RI(m) = p_0 + \sum_{i=1}^2 p_i x_i + \sum_{i=1}^2 p_{ii} x_i^2 + \sum_{i < j} p_{ij} x_i x_j + \epsilon$$

3. KNN imputation: The KNN missing value imputation imputes the missing value of an attribute based on the concept of feature similarity. The missing value is replaced by finding the K closest neighbours.
4. Polynomial Regression: In this research study, polynomial regression is used for imputation of missing values. This method of missing value imputation is used when two variables are related in a nonlinear fashion. The general form of polynomial regression is as follows:

$$y = a + b_1 x + b_2 x^2 + \dots + b_n x^n$$

The main features of using polynomial regression for missing value imputation are as follows:

- (a) It can fit wider range of functions.
- (b) It can fit large range of curvatures.
- (c) It provides a more accurate relationship between two variables.

4.2.2 Standardization

The standardisation techniques used in this experiment are as follows:

1. MixMax scaler (MMS): The MinMax Scaler technique changes each value within a range of 0 and 1. Having column c , the function can be defined as follows:

$$diff[c] = \frac{(diff[c] - diff[c].min())}{diff[c].max() - diff[c].min()}$$

2. Standard Scaler (SS): It standardises an attribute by taking out the mean and then changing it to unit variance.
3. Robust Scaler (RS): Robust scaler algorithms scale attributes that are robust to outliers. It uses the interquartile.

4.2.3 PIMA to Image Dataset

Deep learning models in general and CNN in particular, is superior to conventional machine learning methods in several applications. Applying current CNN models which are made for 2D data, such as images to the PIMA diabetes dataset, which is made up of numerical values, presents a problem. Existing methods use one dimensional CNN models that are specially designed for this dataset. To facilitate the use of well-established CNN models for feature extraction and subsequent diabetes prediction, this work suggests transforming the raw PIMA data. This method seeks to enhance diabetes detection by utilizing deep learning capabilities.

4.2.3.1 Original PIMA Dataset

The NKDDKD originally provided the PIMA dataset which has been used to create machine learning models for early diabetes detection. The dataset comprises nine parameters, namely: age, outcome, skin thickness, blood pressure, glucose concentration, insulin level, body mass index (BMI), diabetes pedigree function, and the number of pregnancies. The collection has 768 entries in total. The dependent variable we are trying to forecast is the "Outcome" attribute. A result of 1 indicates the presence of diabetes, whereas a value of 0 indicates no diabetes. 500 entries have a value of 0 (non-diabetic), according to the analysis of the "Outcome" attribute, while 268 samples with a value of 1 (diabetic). The creation of a universal machine learning model that can diagnose Type I and Type II diabetes is hampered by several issues. Using datasets with inadequate records might be problematic as it can produce erroneous findings. Furthermore, several research uses the PIMA dataset without performing necessary preprocessing procedures like normalization. This may impair the models' accuracy by introducing problems such as outliers, overfitting, and underfitting. Moreover, several research works use restricted machine learning methods for diagnosis and fail to deal with missing values in the data. The limited use of feature extraction algorithms is another drawback. The procedure of extracting features might be greatly enhanced by automatic deep feature extraction. One of the biggest limitations in applying a deep learning model like Resnet50 is not having a large enough dataset to train the model correctly. The study intends to modify this limitation by creating an image-based dataset from PIMA.

4.2.3.2 PIMA Image Dataset

The methodology for converting the PIMA dataset into an image dataset is given in Figure 4.1. As shown in the flowchart, after performing the necessary pre-processing including normalization on the textual dataset, the most pertinent characteristics from the numerical data are extracted using Spearman's correlation coefficient. After data normalization, the bounds of these features are modified for the numeric-to-image conversion step. This approach, which is nonparametric and dependent on correlation, measures the statistical dependency of ranking between two variables. It examines the level of correlation between variables and is represented by ρ . Pearson's Coefficient is seen to be a better choice for feature selection when variables show linear connections. On the other hand, feature selection uses Spearman's Rank Coefficient when variables show monotonic connections. To get Spearman's Rank Coefficient, use this formula:

$$\rho = \frac{6 \sum d^2}{n(n^2 - 1)}$$

The idea of calculating the brightness of a particular area (cell) in the image based on the amplitude of each sample is used in the process of converting PIMA data to images. Each sample in the PIMA dataset is represented as an 8×8 picture structure.

1. The relevant feature value's amplitude determines the colour of each cell in the first row of an 8x8 image.
2. The remaining 7 rows of the image are filled with amplitude values by arranging different pieces or features farther apart while grouping similar elements together; it is possible to make collective use of nearby elements.
3. Since all the data were previously normalized, the values of each feature fall between the range of 0 and 1. By multiplying each feature value by 255, pictures with cells ranging in brightness from 0 to (255 or max value in the sample record) are produced.
4. For each of the 768 records, a total of 7 samples are generated using augmentation techniques such as reflection, rotation and translation etc.
5. The final dataset is generated with 5376 total images.

Figure 4.2 describes the complete methodology for Conversion of Text Pima dataset into Image dataset as well as the use of augmentation techniques:

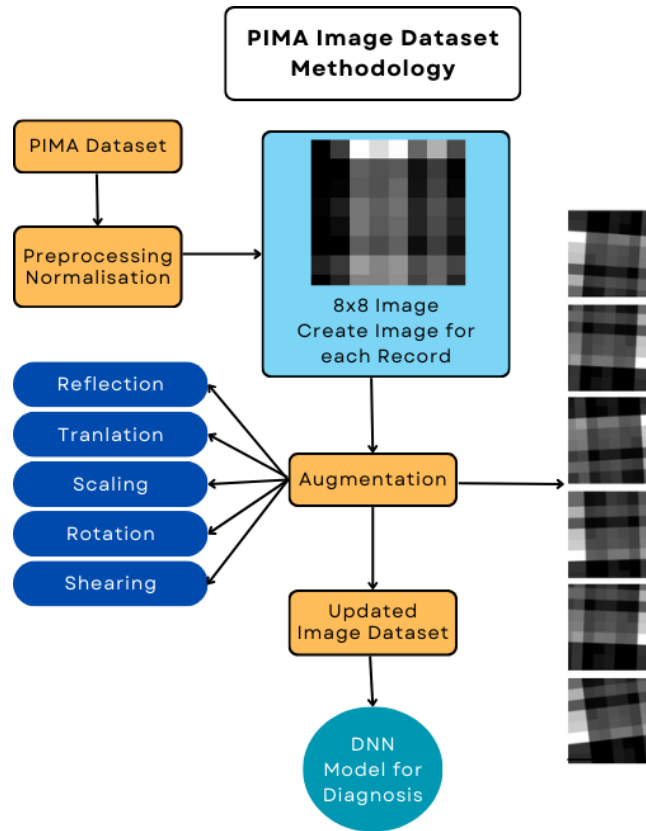


FIGURE 4.2: Methodology for Conversion of Text Pima dataset into Image dataset

Additionally, the application of data augmentation techniques improves the classification accuracy of deep CNN models by increasing the dataset size. From the PIMA dataset containing 768 records, a total of 768 images were obtained. Then the augmentation technique encompassing reflection, rotation, translation, scaling and shearing was employed to generate 7 samples on each sample. The final dataset contains a total of 5376 images. Table 4.1 provides the details about the generated dataset.

TABLE 4.1: Image Dataset Division

	Test (a)	Train (b)	Val (c)	
Yes	190	1504	190	
No	344	2804	344	
Total	534	4308	534	$a+b+c = 5376$

4.2.3.3 DNN-based Model using the Image Dataset

Widely utilized for image classification applications, ResNet50 and VGG16 are robust Convolutional Neural Network (CNN) architectures. ResNet50 has a deeper architecture

making the task of capturing more complex characteristics and relationships in the dataset of images possible. This is advantageous for challenging classification jobs. In comparison to ResNet50, VGG16 has a simpler architecture which runs more quickly and uses less processing power. VGG16 is a dependable baseline option since it is a well-known architecture with a solid reputation in image classification. In this study both these DNN models were used to work with the newly created image dataset on PIMA dataset. The detailed model information for ResNet50 and VGG16 is given in Figure 4.3 and Figure 4.4 respectively.

Layer (type)	Output Shape	Param #
resnet50 (Functional)	(None, 7, 7, 2048)	23587712
dropout_4 (Dropout)	(None, 7, 7, 2048)	0
flatten_2 (Flatten)	(None, 100352)	0
dropout_5 (Dropout)	(None, 100352)	0
dense_2 (Dense)	(None, 1)	100353
=====		
Total params: 23,688,065		
Trainable params: 100,353		
Non-trainable params: 23,587,712		

FIGURE 4.3: Resnet50 Model

Layer (type)	Output Shape	Param #
vgg16 (Functional)	(None, 7, 7, 512)	14714688
dropout (Dropout)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
dropout_1 (Dropout)	(None, 25088)	0
dense (Dense)	(None, 1)	25089
=====		
Total params: 14739777 (56.23 MB)		
Trainable params: 25089 (98.00 KB)		
Non-trainable params: 14714688 (56.13 MB)		

FIGURE 4.4: VGG16 Model

Resnet50 and VGG16 were trained with a total of 120 epochs on pre-trained ImageNet weights. The confusion matrix for resnet50 and VGG16 are shown in Figure 4.5 and Figure 4.6 respectively.

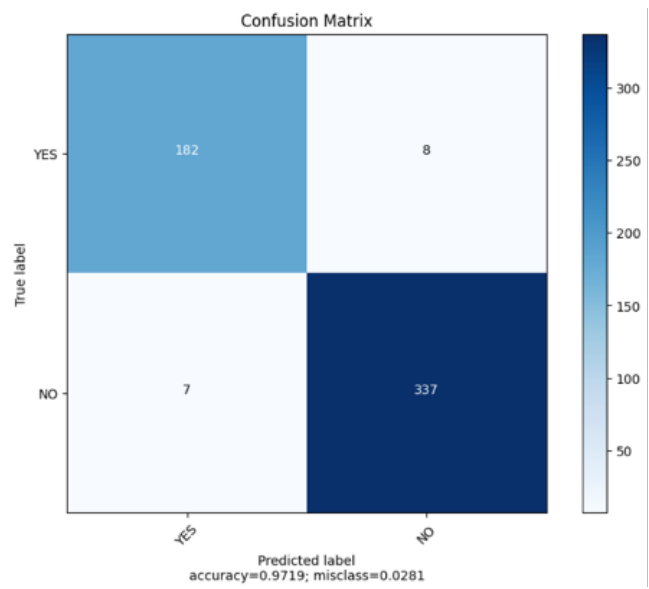


FIGURE 4.5: Confusion MatrixResnet50 Model

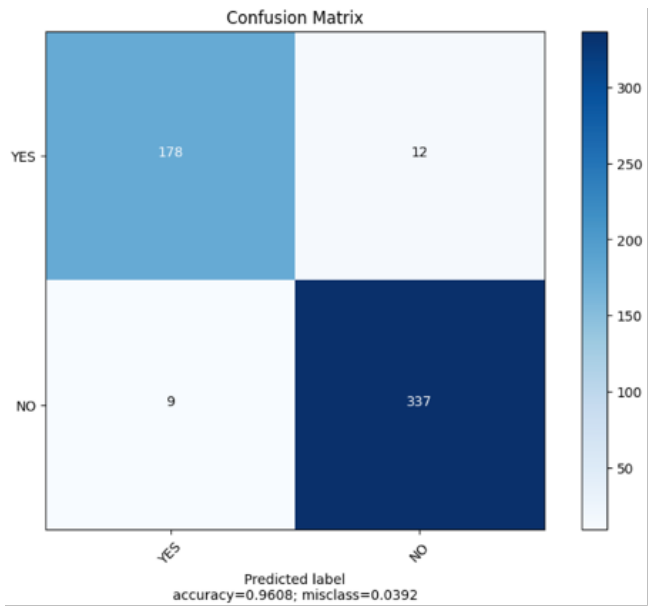


FIGURE 4.6: Confusion Matrix VGG16 Model

Table 4.2 and Table 4.3 present a comparative analysis of the performance of Vgg16, ResNet50, and various research works in the field.

TABLE 4.2: Performance Comparison ReNet50 and VGG16

Measure	VGG16	Resnet50
Sensitivity	0.9368	0.9579
Specificity	0.9738	0.9797
Precision	0.9519	0.9630
NPV	0.9654	0.9768
FPR	0.0262	0.0203
FDR	0.0481	0.0370
FNV	0.0632	0.0421
Accuracy	0.9607	0.9719
F1 Score	0.9443	0.9604

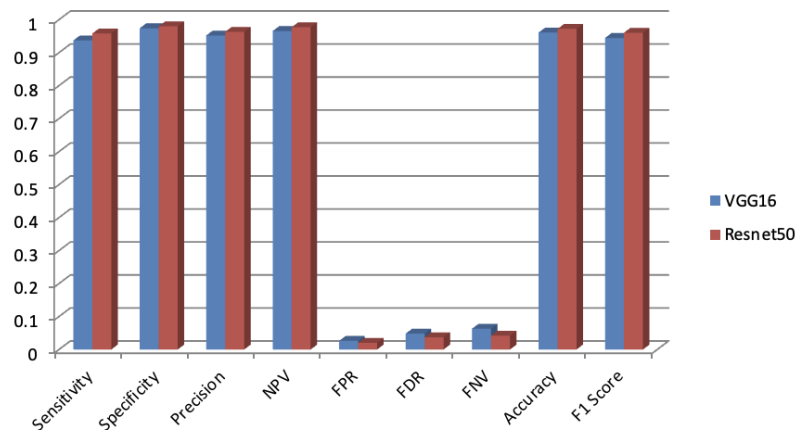


FIGURE 4.7: Performance Chart VGG16 and Resnet50

With accuracy ratings of 96.07% and 97.19% for VGG16 and ResNet50 respectively, it is evident from the Table 4.3 and Figure 4.7 that the VGG16 and ResNet50 models using PIMA image dataset generate the promising results. The method's outcomes indicate that converting diabetic data into an image dataset is a helpful tactic, as seen by the approach's effective classification using VGG16 and ResNet50 models.

TABLE 4.3: Comparative analysis with other DNN-based works

Ref	Technique	Accuracy
[154]	ANN	92.00
[70]	SA-DNN	86.26
[155]	LSTM-CNN	95.7%
[69]	CNN-SAE	92.31%
Proposed method	VGG16 PIMA Image Dataset	96.07%
Proposed method	ResNet50 PIMA Image Dataset	97.19%

4.3 Conclusion

A substantial body of research over the past few decades highlights the value of deep learning for early diagnosis of Diabetes Mellitus (DM). However, the accuracy of deep learning algorithms can be significantly improved through data pre-processing techniques that address issues like overfitting, underfitting, redundancy, missing values, and irrelevant features. This study proposes a methodology that leverages data pre-processing to improve the performance of DL models for DM diagnosis. This methodology has the potential to be applied to a wide range of numerical datasets. While deep learning has reduced the reliance on specific feature selection techniques, the importance of well-designed model architectures remains paramount. This work demonstrates a novel approach that supports the application of deep and complex architectures for numerical data analysis. However in spite of these advancement in the field of AI, Challenges are still there. The main issue is with the size of the dataset. Most of the research was carried out on PIMA dataset which contains limited number of data records and is not suitable for training of deep learning models. The PIMA dataset contains binary data in numeric form and lacks the opportunity of applying data augmentation. Although image data analysis might require additional pre-processing steps compared to studies using raw data, this approach presents exciting opportunities to improve DM prediction capabilities. This is facilitated by the ability of Convolutional Neural Network (CNN) models to adapt to numerical inputs. Furthermore, the integration of data augmentation techniques becomes more straightforward when dealing with diabetes-related images. This allows for the creation of a more robust training dataset and potentially leads to more accurate and generalizable models for DM diagnosis.

Chapter 5

Conclusion and Future Directions

5.1 Conclusion

This study used the PIMA dataset in its raw form to train multiple machine learning models and log the effectiveness of different ML algorithms for diabetes mellitus diagnosis. This allows the results of the proposed model to be compared with other available research studies conducted in the field. Following this, various pre-processing techniques like standardization techniques, normalization techniques, MVI techniques and FS and feature importance methods to eliminate various anomalies present in PIMA dataset and then train various machine language models to get better results have been utilized. The outcomes were contrasted with the earlier findings. Finally, the textual Pima dataset is transformed into image dataset with 768 images (one image per record). After conversion into image dataset, various data augmentation techniques have been applied on image dataset to get more images. Using various data augmentation techniques like rotation, transformation, scaling etc., the number of images generated was more than 5000. After conversion into image dataset, various data augmentation techniques have been applied on image dataset to get more images. The trained model was evaluated against the output of other researchers as well as with findings from earlier studies utilizing an image dataset which signified that the proposed model is more promising in diagnosis of DM.

From the comprehensive overview of DM, its prevalence, and the significance of early diagnosis, One way to classify diabetes mellitus (DM) is as a chronic illness marked by high blood sugar levels. This is because the body is unable to generate or use insulin, a hormone that is essential for controlling blood sugar levels. Diabetes comes in several

forms, such as Type 1, Type 2, and gestational diabetes, having consequent symptoms associated with each type.

The importance of early diagnosis is crucial, as it plays a critical role in preventing serious consequences such as blindness, renal disease, cardiac issues etc. There are limitations of current diagnostic methods, that often rely on blood glucose tests, HbA1c tests, and urine tests, and more efficient and trustworthy techniques for early detection are desperately needed.

Understanding the possibilities of deep learning (DL) and machine learning (ML) methods in aiding diabetes diagnosis, their ability to analyze vast datasets and uncover hidden patterns often missed by traditional methods, a comprehensive literature review, examining previous research and models, focusing on these techniques in diabetes prediction. The review showcases the advancements and successes achieved in this field, paving the way for future innovations.

There is a plethora of complexities associated with designing and developing models for DM diagnosis, especially the challenges posed by the disease's nature and the availability of data. Also, factors such as the non-specific nature of diabetes symptoms, and the limited size and quality of existing datasets need to be addressed.

The study established that Random Forests, Gradient Boost, and Logistic Reasoning classifiers performed better and should be considered for future research, incorporating all significant parameters that may limit classifier performance. The DL-based techniques require a larger dataset, a DNN shall be preferred only when cross-validation is integrated.

Deep learning techniques have shown significant promise in diagnosis of various diseases at early stage and the same techniques can be applied to DM early diagnosis. From the literature review, various advantages of deep learning over machine learning models were detected which include handling of large datasets, uncovering complex hidden patterns and accurate diagnosis. From the studies it was observed that deep learning models like CNN and RNN outperforms conventional machine learning techniques in terms of forecast accuracy and dependability of prediction. Further, the use of Deep Learning with image datasets like medical imaging provides better approach to diagnosis of diabetes mellitus. The application of data augmentation, normalization, dimensional reduction and standardization has further improved the robustness and generalization of deep learning models. However, in spite of these advancement in the field of AI, Challenges are still there. The main issue is with the size of the dataset. Most of the research was carried out on PIMA dataset which contains limited number of data records and is not suitable for training of deep learning models. The PIMA dataset contains binary data in numeric form and lacks the opportunity of applying data augmentation.

A review of the literature reveals that many researchers use the PIMA dataset without applying pre-processing techniques such as normalization. Consequently, their findings are affected by outliers, overfitting, underfitting, and other anomalies. Additionally, some studies employ a limited range of machine learning algorithms for diagnosing diabetes and do not address missing values. There are also instances where the full potential of feature extraction is not utilized. Certain studies overlook the importance of all dataset attributes, particularly those like body size, height, and BMI, which significantly aid in the determination of DM diagnosis. This oversight negatively impacts the performance of classifiers. The results of various studies indicate that Random Forests provide the good accuracy which is followed by Gradient Boosting and Logistic Regression. For the early detection of diabetes mellitus, these algorithms should be considered, along with the incorporation of the aforementioned insights, to improve output and performance.

The role of pre-processing techniques in improving the accuracy and reliability of diabetes diagnosis models is very crucial. Supervised learning methods rely on labeled datasets to train models, while unsupervised learning methods discover patterns in unlabeled data. Some commonly used supervised learning techniques such as DT, ANN, and SVM. DL techniques, particularly DBN and RNN have become more well-known recently as a result of their capacity to manage intricate data patterns and big datasets. These techniques can be applied to diabetes diagnosis, showcasing their potential to outperform ML methods.

To demonstrate the impact of pre-processing techniques, several experiments on a publicly available diabetes dataset, the PIMA dataset, without any pre-processing were conducted. It provided a baseline for comparison with the results obtained after applying various pre-processing techniques. The PIMA dataset, while valuable, presents certain challenges, including limited data records, binary data in numeric form, and a lack of opportunity for data augmentation. This underscores the importance of effective pre-processing to address these limitations.

The pre-processing steps applied to the PIMA dataset, including data cleansing, data transformation, FS, and feature engineering aimed to enhance the quality and structure of the data, making it more suitable for training and evaluating diagnosis models. Various techniques such as managing missing values, handling outliers, normalization, encoding categorical variables, and feature selection models help overcome many limitations found in the PIMA dataset.

The results of the experiments with and without pre-processing, demonstrate the significant improvement in model performance achieved by applying the pre-processing techniques. Based on these methods, a DNN model for diabetes diagnosis highlights

the critical role of pre-processing in optimizing data for accurate and reliable diabetes diagnosis.

A ground-breaking approach to address the limitations of conventional convolutional neural network (CNN) models when applied to numerical medical datasets was proposed. The PIMA Indians Diabetes Database, while valuable, presents challenges due to its numerical nature, hindering the effectiveness of CNN models. To overcome this obstacle, a novel technique for transforming numerical data into visual representations based on feature relevance was proposed. This conversion opens the door for utilizing powerful CNN models for early diabetes diagnosis.

Further, deep learning, a highly effective technique for early diagnosis of diabetes mellitus, can be enhanced by pre-processing techniques that eliminate anomalies such as over-fitting, under-fitting, redundancy, missing values, and non-significant features. This optimized dataset significantly improves the efficiency of deep learning algorithms for diagnosing the disease. By transforming numerical data into visual representations, the power of CNN models can be leveraged to extract patterns and identify hidden relationships in the data, ultimately leading to more accurate and timely diagnosis. This innovative approach paves the way for a new era in diabetes prediction and management.

5.2 Future Directions

The research in this field has to go a long way. The application of DL-based models on image dataset is now a pragmatic idea but many of the models used today only take into account clinical or blood test data. In future, the researchers may focus on incorporating Wearable devices data on blood glucose, heart rate variability, sleep patterns, and activity levels might give important insights into a patient's health. Also, knowing a person's genetic susceptibility to diabetes can help with risk assessment and early identification. finally, the information for diagnosis and problem prediction may be provided by retinal scans, pictures of foot ulcers, and other medical imaging studies.

The significance of developing sophisticated pre-processing techniques, such as novel data augmentation or regularization methodologies, is profound. While the current work focused on a methodical comparison of established approaches to optimize the PIMA dataset for comparative analysis of ML methods and the particular image transformation and deep learning application, laying the essential foundation for the innovative contribution, future research shall focus on examining and implementing advanced pre-processing strategies to overcome the limitations of the dataset used in ML techniques

and DNN's reported limitations, potentially enhancing the model's robustness and performance.

In the context of DM detection using deep learning, combining DNA data with the PIMA dataset may be a viable next step. The most straightforward method is fusing pertinent DNA sequence data with the PIMA dataset's attributes (such as glucose levels and BMI). This can entail taking particular genetic markers or variations that are known to be linked to the risk of diabetes and adding them to the PIMA data as supplementary columns. This merged dataset might then be used to train deep learning models.

Future research might concentrate on feature engineering rather than directly utilizing raw DNA sequences. This entails identifying significant characteristics in the DNA data, like:

1. Polygenic Risk Scores (PRS): Determine scores that represent the total impact of several genetic variations on a person's risk of diabetes.
2. Gene Expression Levels: Determine how active certain genes linked to insulin sensitivity or glucose metabolism are.
3. Epigenetic Markers: Include details regarding histone modifications or DNA methylation, which might affect how genes are expressed.

Multi-modal deep learning architectures may be used in more complex methods. These models are made to process several kinds of data at once. The clinical characteristics of the PIMA dataset might be processed by a single network branch. The characteristics generated from DNA might be processed by another branch. A final forecast may then be created by fusing the results of these branches.

Preventing the development of full-blown diabetes requires early diagnosis of pre-diabetes. Subtle alterations in blood indicators or other data points that can refer to the disease's early symptoms can be detected using DL models.

However, deep learning's high computing cost, environmental effect, and lack of intrinsic sustainability are some of its major issues. To overcome this, a multifaceted strategy emphasizing hardware optimization, algorithmic efficiency, and investigating alternative paradigms is needed.

Several approaches that concentrate on algorithmic efficiency are essential to addressing deep learning's high computational cost and environmental effect. Pruning, quantization, and knowledge distillation are examples of model compression approaches that may drastically minimize the computing footprint of deep learning models, allowing for quicker inference and less energy usage. To create models that need fewer parameters

and processes, research on effective designs like sparse networks and attention mechanisms is also crucial. A more sustainable approach to deep learning may also be achieved by using more effective optimization techniques during training and adopting adaptive computing, in which models dynamically modify their resource use based on input complexity.

Overcoming the difficulties of deep learning requires not only algorithmic advancements but also hardware optimization and the investigation of alternative paradigms. Deep learning activities may be accelerated and energy efficiency increased by utilizing specialized hardware such as GPUs, TPUs, and neuromorphic devices. By using edge computing to deploy models on edge devices, data transport and the related energy expenses are decreased. Investigating paradigms for computing inspired by the brain, such as spiking neural networks, has the potential to significantly reduce power use.

It is pertinent to mention that the cost of computation and time estimation are important aspects. Deep learning models' applicability and scalability are significantly impacted by these variables, especially in settings with limited resources. Future studies can provide deeper learning systems that are more effective and economical for a range of applications by taking these extra performance metrics into account.

Appendix A

An Appendix

TABLE A.1: Definitions

Description	Definition
ACCuracy	Accuracy is a metric for evaluating classification models.
AdaBoosted Decision Trees	Using AdaBoost to improve performance in decision trees.
AdaBoostRegressor	Using AdaBoost to improve performance in regression.
Adaptive Boosting	A statistical classification meta-algorithm that can be used in conjunction with many other types of learning algorithms to improve performance.
Area Under the (ROC) Curve	Probability of confidence in a model to accurately predict positive outcomes for actual positive instances
Artificial Intelligence	The simulation of human intelligence in machines that are programmed to think like humans and mimic their actions.
Artificial Neural Network	A collection of connected computational units or nodes called neurons arranged in multiple computational layers.

AutoEncoder	A type of artificial neural network used to learn efficient codings of unlabeled data (unsupervised learning)		
BackPropagation	A widely used algorithm for training feedforward neural networks.		
Bayesian Network	A probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).		
Bayesian Neural Network	A type of artificial neural network built by introducing random variations into the network either by giving the network's artificial neurons stochastic transfer functions either by giving the network's artificial neurons stochastic transfer functions or by giving them stochastic weights		
Convolutional Networks	Deep	Belief	A type of deep artificial neural network composed of multiple layers of convolutional restricted Boltzmann machines stacked together.
Convolutional Neural Network	Neural	Net-	A class of artificial neural network (ANN) most commonly applied to analyze visual imagery
Convolutional Neural Network	Neural	Net-	A class of artificial neural network (ANN) most commonly applied to analyze visual imagery
False Negative Rate	Proportion of actual positives predicted as negatives		
False Positive Rate	Proportion of actual negatives predicted as positives		
Fully Connected Long Short-Term Memory	A fully connected neural network to combine the spatial information of surrounding stations (see LSTM and FC).		
Fully Convolutional Convolutional Neural Network	A neural network that only performs convolution (and subsampling or upsampling) operations.		
Fully Convolutional Network	A neural network that only performs convolution (and subsampling or upsampling) operations.		

Fully-Connected	Layers where all the inputs from one layer are connected to every activation unit of the next layer.
Gradient Descent	An optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient
k-Nearest Neighbours	A non-parametric supervised learning method used for classification and regression.
Light Gradient-Boosting Machine	Gradient boosting framework that uses tree based learning algorithms, originally developed by Microsoft
Long Short-Term Memory	A recurrent neural network can process not only single data points (such as images) but also entire sequences of data (such as speech or video).
Machine Learning	The study of computer algorithms that can improve automatically through experience and by the use of data.
Mean Absolute Error	Average of the absolute error between the actual and predicted values
Mean Squared Error	Average of the squares of the error between the actual and predicted values
Rectified Linear Unit	An activation function that allow fast and effective training of deep neural architectures on large and complex datasets.
Support Vector Machine	Supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.
True Negative Rate	Proportion of actual negatives that are correctly predicted
True Positive Rate	Proportion of actual positives that are correctly predicted

Bibliography

- [1] Ovass Shafi, S Jahangeer Sidiq, Tawseef Ahmed Teli, and Majid Zaman. A Comparative Study on Various Data Mining Techniques for Early Prediction of Diabetes Mellitus. Technical report, 2021.
- [2] Larissa Carvalho, Ana Vitória Bezerra Silva, Breno José de Alencar Danda, Elaine Cristina Batista Ferreira Freitas, and Moises Thiago de Souza Freitas. Polimorfismos no gene transportador de glicose (GLUT4) associados ao desenvolvimento da Diabetes mellitus tipo 1. *Research, Society and Development*, 11(13):e368111335549, oct 2022. ISSN 2525-3409. doi: 10.33448/rsd-v11i13.35549. URL <https://rsdjournal.org/index.php/rsd/article/view/35549>.
- [3] Masafumi Koga, Ikki Shimizu, Jun Murai, Hiroshi Saito, Soji Kasayama, Tetsuro Kobayashi, Akihisa Imagawa, and Toshiaki Hanafusa. The glycated albumin to HbA1c ratio is elevated in patients with fulminant type 1 diabetes mellitus with onset during pregnancy. *Journal of Medical Investigation*, 60(1-2):41–45, 2013. ISSN 13496867. doi: 10.2152/jmi.60.41.
- [4] Hafiz Farooq Ahmad, Hamid Mukhtar, Hesham Alaqail, Mohamed Seliaman, and Abdulaziz Alhumam. Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Applied Sciences (Switzerland)*, 11(3):1–18, feb 2021. ISSN 20763417. doi: 10.3390/app11031173.
- [5] Devi Kalyan Karumanchi, Elizabeth R. Gaillard, and James Dillon. Early Diagnosis of Diabetes through the Eye. *Photochemistry and Photobiology*, 91(6):1497–1504, nov 2015. ISSN 17511097. doi: 10.1111/php.12524.
- [6] Andrzej Grzybowski, Piotr Brona, Gilbert Lim, Paisan Ruamviboonsuk, Gavin S.W. Tan, Michael Abramoff, and Daniel S.W. Ting. Artificial intelligence for diabetic retinopathy screening: a review, mar 2020. ISSN 14765454.

- [7] Diego Micael Barreto Andrade, Roseanne Montargil Rocha, and Ícaro José Santos Ribeiro. Depressive symptoms among older adults with diabetes mellitus: a cross-sectional study. *Sao Paulo Medical Journal*, oct 2022. ISSN 1806-9460. doi: 10.1590/1516-3180.2021.0771.r5.09082022. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-31802022005025201&tlng=en.
- [8] Noha E. El-Attar, Bossy M. Moustafa, and Wael A. Awad. Deep learning model to detect diabetes mellitus based on dna sequence. *Intelligent Automation and Soft Computing*, 31(1):325–338, 2022. ISSN 2326005X. doi: 10.32604/IASC.2022.019970.
- [9] Adnan Hashim, Zukhruf Masood, and Fareeha Amjad. Frequency of restless legs syndrome in patients with diabetes mellitus and hypertension. Technical Report 3. URL <https://www.researchgate.net/publication/364164864>.
- [10] Mudasir Ashraf, Majid Zaman, Muheet Ahmed Butt, and Muheet S Ahmed Jahangeer Sidiq. Impact of performance analysis of varied subjects on overall result: An empirical discourse of educational data mining View project SAMIAH NASTI View project Knowledge Discovery in Academia: A Survey on Related Literature. *International Journal of Advanced Research in Computer Science*, 8(1). ISSN 0976-5697. URL www.ijarcs.info.
- [11] Jyotismita Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, 34(6):3204–3225, 2022. ISSN 22131248. doi: 10.1016/j.jksuci.2020.06.013. URL <https://doi.org/10.1016/j.jksuci.2020.06.013>.
- [12] Kuang Ming Kuo, Paul Talley, Yu Hsi Kao, and Chi Hsien Huang. A multi-class classification model for supporting the diagnosis of type II diabetes mellitus. *PeerJ*, 8, 2020. ISSN 21678359. doi: 10.7717/peerj.9920.
- [13] Nasloon Ali, Wasif Khan, Amir Ahmad, Mohammad Mehedy Masud, Hiba Adam, and Luai A. Ahmed. Predictive Modeling for the Diagnosis of Gestational Diabetes Mellitus Using Epidemiological Data in the United Arab Emirates. *Information*, 13(10):485, oct 2022. ISSN 2078-2489. doi: 10.3390/info13100485. URL <https://www.mdpi.com/2078-2489/13/10/485>.
- [14] Juan Juan, Yiyang Sun, Yumei Wei, Shuang Wang, Geng Song, Jie Yan, Pengxiang Zhou, and Huixia Yang. Progression to type 2 diabetes mellitus after gestational

- diabetes mellitus diagnosed by IADPSG criteria: Systematic review and meta-analysis. *Frontiers in Endocrinology*, 13, oct 2022. ISSN 1664-2392. doi: 10.3389/fendo.2022.1012244. URL <https://www.frontiersin.org/articles/10.3389/fendo.2022.1012244/full>.
- [15] Kameran H Ismail, Yahya Adil Hasan, and Kameran Hassan Ismail. Prevalence of Type Two Diabetes Mellitus and Risk Factors in Erbil City among Adult Population Yahya Adil et al / Prevalence of Type Two Diabetes Mellitus and Risk Factors in Erbil City among Adult Population Prevalence of Type Two Diabetes Mellitus and Risk Factors in Erbil City among Adult Population. doi: 10.14704/nq.2022.20.10.NQ55221. URL www.neuroquantology.com.
- [16] Fareeha Anwar, Qurat-Ul-Ain, Muhammad Yasir Ejaz, and Amir Mosavi. A comparative analysis on diagnosis of diabetes mellitus using different approaches – A survey. *Informatics in Medicine Unlocked*, 21, jan 2020. ISSN 23529148. doi: 10.1016/j.imu.2020.100482.
- [17] Veena Mayya, Sowmya Kamath S, and Uma Kulkarni. Automated microaneurysms detection for early diagnosis of diabetic retinopathy: A Comprehensive review. *Computer Methods and Programs in Biomedicine Update*, 1:100013, 2021. ISSN 26669900. doi: 10.1016/j.cmpbup.2021.100013.
- [18] Olusola Olabanjo, Manuel Mazzara, and Ashiribo Wusu. Deep Unsupervised Machine Learning for Early Diabetes Risk Prediction using Ensemble Feature Selection and Deep Belief Neural Networks Prediction of Twitter Message Deletion View project AutoReq View project Deep Unsupervised Machine Learning for Early Diabetes Risk Prediction using Ensemble Feature Selection and Deep Belief Neural Networks. 2023. doi: 10.20944/preprints202301.0208.v1. URL www.preprints.org.
- [19] Seyed Ataaldin Mahmoudinejad Dezfuli, Seyedeh Razieh Mahmoudinejad Dezfuli, Seyed Vafaaldin Mahmoudinejad Dezfuli, and Younes Kiani. Early Diagnosis of Diabetes Mellitus Using Data Mining and Classification Techniques. *Jundishapur Journal of Chronic Disease Care*, 8(3), jul 2019. ISSN 2322-3758. doi: 10.5812/jjcdc.94173.
- [20] Muqiu Zhang and Huixia Yang. Perspectives from metabolomics in the early diagnosis and prognosis of gestational diabetes mellitus. *Frontiers in Endocrinology*, 13, sep 2022. doi: 10.3389/fendo.2022.967191.

- [21] Onur SEVLİ. Diyabet hastalığının farklı sınıflandırıcılar kullanılarak teşhisi. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, may 2022. ISSN 1300-1884. doi: 10.17341/gazimmfd.880750.
- [22] Kareem Arab, Zied Bouida, and Mohamed Ibnkahla. Artificial Intelligence for Diabetes Mellitus Type II : Forecasting and Anomaly Detection. *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2019. doi: 10.1109/WCNC.2019.8885802.
- [23] Derara Duba Rufo, Taye Girma Debelee, Achim Ibenthal, and Worku Gachena Negera. Diagnosis of diabetes mellitus using gradient boosting machine (Lightgbm). *Diagnostics*, 11(9), sep 2021. ISSN 20754418. doi: 10.3390/diagnostics11091714.
- [24] Chollette C. Olisah, Lyndon Smith, and Melvyn Smith. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, jun 2022. ISSN 18727565. doi: 10.1016/j.cmpb.2022.106773.
- [25] G. Swapna, R. Vinayakumar, and K. P. Soman. Diabetes detection using deep learning algorithms. *ICT Express*, 4(4):243–246, 2018. ISSN 24059595. doi: 10.1016/j.ict.2018.10.005.
- [26] Hyun Shik Son. Early Diagnosis of Diabetes Mellitus. Technical report.
- [27] Ratna Patil, Sharvari Tamane, Shitalkumar Adhar Rawandale, and Kanishk Patil. A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus. *International Journal of Electrical and Computer Engineering*, 12(1):524–533, feb 2022. ISSN 20888708. doi: 10.11591/ijece.v12i1.pp524-533.
- [28] B. Shamreen Ahamed, Meenakshi S. Arya, and Auxilia Osvin V. Nancy. Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Over-sampling and Feature Augmentation. *Advances in Human-Computer Interaction*, 2022, 2022. ISSN 16875907. doi: 10.1155/2022/9220560.
- [29] R Karthikeyan, P Geetha, and E Ramaraj. THE RULE-BASED MULTI-CLASS CLASSIFICATION MODEL PREDICTS EARLY DIABETES USING SUPERVISED MACHINE LEARNING TECHNIQUES. 2022. ISSN 1005-3026. URL <https://dbdxxb.cn/>.

- [30] Ployphan Sornsuwit. ENHANCE WEAK LEARNER MODEL OF ADABOOST (EWDM) FOR DIABETES MELLITUS CLASSIFICATION. *International Journal of Innovative Computing, Information and Control*, 18(4):1117–1132, aug 2022. ISSN 13494198. doi: 10.24507/ijicic.18.04.1117.
- [31] Nisreen Sulayman. Predicting Type 2 Diabetes Mellitus using Machine Learning Algorithms Digital Image Processing View project. Technical report. URL <https://www.researchgate.net/publication/366634353>.
- [32] Jobeda Jamal Khanam and Simon Y. Foo. A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4):432–439, dec 2021. ISSN 24059595. doi: 10.1016/j.ict.2021.02.004.
- [33] Zahura Zaman, Md. Ashrak Al Arif Shohas, Mahedi Hasan Bijoy, Meherab Hos-sain, and Shakawat Al Sakib. Assessing Machine Learning Methods for Predicting Diabetes among Pregnant Women. *International Journal of Advancement in Life Sciences Research*, 05(01):29–34, 2022. doi: 10.31632/ijalsr.2022.v05i01.005.
- [34] Ammar Armghan, Jaganathan Logeshwaran, S. M. Sutharshan, Khaled Aliqab, Meshari Alsharari, and Shobhit K. Patel. Design of biosensor for synchro-nized identification of diabetes using deep learning. *Results in Engineering*, 20 (September):101382, 2023. ISSN 25901230. doi: 10.1016/j.rineng.2023.101382. URL <https://doi.org/10.1016/j.rineng.2023.101382>.
- [35] Md Shahriare Satu, Syeda Tanjila Atik, Mohammad Ali Moni, and Syeda Tanjila Atik. A Novel Hybrid Machine Learning Model To Predict Diabetes Mellitus Modelling and formal analysis of bone remodelling View project Comorbidities of Thyroid Cancer View project A Novel Hybrid Machine Learning Model To Predict Diabetes Mellitus. Technical report, 2019. URL <https://www.researchgate.net/publication/335727823>.
- [36] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, and Xiaoyi Wang. Informatics in Medicine Unlocked Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, 10(August 2017):100–107, 2018. ISSN 2352-9148. doi: 10.1016/j.imu.2017.12.006. URL <https://doi.org/10.1016/j.imu.2017.12.006>.
- [37] Amin Ul Haq, Jian Ping Li, Jalaluddin Khan, Muhammad Hammad Memon, Shah Nazir, Sultan Ahmad, Ghufraan Ahmad Khan, and Amjad Ali. Intelligent machine learning approach for effective recognition of diabetes in e-healthcare

- using clinical data. *Sensors (Switzerland)*, 20(9), may 2020. ISSN 14248220. doi: 10.3390/s20092649.
- [38] Aleena Farooq, Muhammad Kamran Abid, Wasif Akbar, Hafiz Humza, and Naeem Aslam. Type-II Diabetes Prediction by using Classification and Novel based Method (AWOD). *Journal of Computing Biomedical Informatics*, 4 (01):152–174, dec 2022. ISSN 2710-1614. doi: 10.56979/401/2022/110. URL <https://jcbi.org/index.php/Main/article/view/110>.
- [39] Gabriel Spiridon, Anca Sarbu, Dorin Carstoiu, and Florian Ion. Computerised decision system for diabetes mellitus and associated complications - CODES. *2018 IEEE International Conference on Automation, Quality and Testing, Robotics, AQTR 2018 - THETA 21st Edition, Proceedings*, pages 1–4, 2018. doi: 10.1109/AQTR.2018.8402772.
- [40] Xi Wang, Jianlin Yu, Zhenhao Li, Jianzheng Hu, Chenglin Sun, Lili He, Hongtao Bai, and Research Article. A Revised Adaptive Network-based Fuzzy Inference System Combined with Neural Network to Predict Diabetes. 2022. doi: 10.21203/rs.3.rs-2388120/v1. URL <https://doi.org/10.21203/rs.3.rs-2388120/v1>.
- [41] Shamim Reza, Ruhul Amin, Rubia Yasmin, Woomme Kulsum, and Sabba Ruhi. Heliyon Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. *Heliyon*, 10(2):e24536, 2024. ISSN 2405-8440. doi: 10.1016/j.heliyon.2024.e24536. URL <https://doi.org/10.1016/j.heliyon.2024.e24536>.
- [42] Kiran D Yesugade, Harshada V Ankam, Anushka A Urunkar, Poonam D Dede, and Sonal S Kale. MACHINE LEARNING BASED WEB APPLICATION FOR DIABETES PREDICTION. Technical report, 2022. URL www.jetir.org/h488.
- [43] Fayzeh Abdulkareem Jaber and Joy Winston James. Early Prediction of Diabetic Using Data Mining. *SN Computer Science*, 4(2), jan 2023. ISSN 26618907. doi: 10.1007/s42979-022-01594-z.
- [44] Leila Ismail and Huned Materwala. IDMPF: intelligent diabetes mellitus prediction framework using machine learning. *Applied Computing and Informatics*, 2021. ISSN 2634-1964. doi: 10.1108/aci-10-2020-0094.
- [45] Munish Khanna, Law Kumar Singh, Shankar Thawkar, and Mayur Goyal. Deep learning based computer-aided automatic prediction and grading system for diabetic retinopathy. *Multimedia Tools and Applications*, mar 2023. ISSN 1380-7501.

- doi: 10.1007/s11042-023-14970-5. URL <https://link.springer.com/10.1007/s11042-023-14970-5>.
- [46] Sun Ok Song, Jae Seung Yun, Seung Hyun Ko, Yu Bae Ahn, Bo Yeon Kim, Chul Hee Kim, Ja Young Jeon, Dae Jung Kim, Da Hae Seo, So Hun Kim, Jung Hyun Noh, Da Young Lee, Kyung Soo Kim, and Soo Kyung Kim. Prevalence and clinical characteristics of fulminant type 1 diabetes mellitus in Korean adults: A multi-institutional joint research. *Journal of Diabetes Investigation*, 13(1):47–53, jan 2022. ISSN 20401124. doi: 10.1111/jdi.13638.
- [47] Nilarun Mukherjee and Souvik Sengupta. In Search for the Optimal Preprocessing Technique for Deep Learning Based Diabetic Retinopathy Stage Classification from Fundus Images. doi: 10.21203/rs.3.rs-654484/v1. URL <https://doi.org/10.21203/rs.3.rs-654484/v1>.
- [48] Hima Tallam, Daniel C. Elton, Sungwon Lee, Paul Wakim, Perry J. Pickhardt, and Ronald M. Summers. Fully Automated Abdominal CT Biomarkers for Type 2 Diabetes Using Deep Learning. *Radiology*, 304(1):85–95, jul 2022. ISSN 15271315. doi: 10.1148/radiol.211914.
- [49] Krishna Babu Ojha and Hritik Verma. Potential effect of Thuja orientalis leaves extract in Scopolamine induced Alzheimer View project Comprehensive Study on Drug Designing and Process Chemistry: Where we are now and What lies ahead View project. doi: 10.37896/YMER22.02/18. URL <https://www.researchgate.net/publication/368291636>.
- [50] Abdulhalim Salim Serafi and Zahir Hussain. Diabetes mellitus: Laboratory diagnosis Pathophysiology of Metabolic Syndrome View project Pathophysiology of ischemic diseases View project. doi: 10.13140/RG.2.2.21901.15845. URL <https://www.researchgate.net/publication/348835661>.
- [51] Prakash D. Effect of cinnamon on diabetes mellitus EFFECT OF SPIRITUAL THERAPY ON QUALITY OF SLEEP View project Effect of sensitization program on prevention of polycystic ovarian syndrome among adolescent girls View project A Study to Evaluate the Effectiveness of Cinnamon on Blood Glucose among Drivers with Type II Diabetes Mellitus at Dhanalakshmi Srinivasan Group of Institutions, Perambalur. *International Journal of Science and Research*, 2018. ISSN 2319-7064. doi: 10.21275/SR20320145202. URL <https://www.researchgate.net/publication/366192336>.

- [52] Abhilash Pati, Manoranjan Parhi, Binod Kumar Pattanayak, Debabrata Singh, Debabrata Samanta, Amit Banerjee, Sajal Biring, and Goutam Kumar Dalapati. Diagnose Diabetic Mellitus Illness Based on IoT Smart Architecture. *Wireless Communications and Mobile Computing*, 2022, 2022. ISSN 15308677. doi: 10.1155/2022/7268571.
- [53] Mritunjay Rai, Tanmoy Maity, Rohit Sharma, and R. K. Yadav. Early detection of foot ulceration in type II diabetic patient using registration method in infrared images and descriptive comparison with deep learning methods. *Journal of Supercomputing*, 78(11):13409–13426, jul 2022. ISSN 15730484. doi: 10.1007/s11227-022-04380-z.
- [54] Osama Shaikh Omar, Zahir Hussain View project Abdulhalim salim Serafi, and Zahir Hussain. Diabetes Mellitus: Laboratory diagnosis. Technical report.
- [55] G. Swapna, K. P. Soman, and R. Vinayakumar. Diabetes Detection Using ECG Signals: An Overview. In *Studies in Big Data*, volume 68, pages 299–327. Springer Science and Business Media Deutschland GmbH, 2020. doi: 10.1007/978-3-030-33966-1_14.
- [56] Hui Wang, Ye Yao, Jieying Zheng, Danhong Peng, Jiansheng Wu, and Jun Wang. Accurate prediction of gestational diabetes mellitus via a novel transformer method. 2023. doi: 10.21203/rs.3.rs-2461259/v1. URL <https://doi.org/10.21203/rs.3.rs-2461259/v1>.
- [57] Wenhao Jiang, Sowmya S Kamath, Yun Liu, fpsyt Copyright, Cheng Wan, Wei Feng, Renyi Ma, Hui Ma, Junjie Wang, Ruochen Huang, Xin Zhang, Mang Jing, Hao Yang, and Haoran Yu. OPEN ACCESS EDITED BY Association between depressive symptoms and diagnosis of diabetes and its complications: A network analysis in electronic health records. Technical report.
- [58] Julia Pastorello, Emanuela Lando, Lorenzo Gervaso, and Nicola Fazio. Association of Diabetes Mellitus and Pancreatic Cancer: Literature Review. *Brazilian Journal of Case Reports*, 2023:29. doi: 10.52600/2763-583X.bjcr.2023.3.2.29-34. URL <https://doi.org/10.52600/2763-583X.bjcr.2023.3.2.29-34>.
- [59] Farsad Zamani Boroujeni, Razieh Asgarnezhad, and Maryam Shekofteh. Improving diagnosis of diabetes mellitus using combination of preprocessing techniques. *Article in Journal of Theoretical and Applied Information Technology*, 15(13), 2017. ISSN 1817-3195. URL <https://www.researchgate.net/publication/318777104>.

- [60] Daria Di Filippo, Amanda Henry, Chloe Bell, Sarah Haynes, Melissa Han Yiin Chang, Justine Darling, and Alec Welsh. A new continuous glucose monitor for the diagnosis of gestational diabetes mellitus: a pilot study. *BMC Pregnancy and Childbirth*, 23(1):186, mar 2023. ISSN 1471-2393. doi: 10.1186/s12884-023-05496-7. URL <https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-023-05496-7>.
- [61] Muhammet Fatih Aslan and Kadir Sabanci. A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data. *Diagnostics*, 13(4), 2023. ISSN 20754418. doi: 10.3390/diagnostics13040796.
- [62] Sarita Simaiya, Rajwinder Kaur, Jasminder Kaur Sandhu, Majed Alsafyani, Roobaea Alroobaea, Deema mohammed Alsekait, Martin Margala, and Prasun Chakrabarti. A novel multistage ensemble approach for prediction and classification of diabetes. *Frontiers in Physiology*, 13, dec 2022. ISSN 1664042X. doi: 10.3389/fphys.2022.1085240.
- [63] Abdulhakim Salum Hassan, I. Malaserene, and A. Anny Leema. Diabetes Mellitus Prediction using Classification Techniques. *International Journal of Innovative Technology and Exploring Engineering*, 9(5):2080–2084, 2020. doi: 10.35940/ijitee.e2692.039520.
- [64] Shaimaa Hameed and Farah Al-Khalidi. DIAGNOSIS OF DIABETES MELLITUS BASED COMBINED OF FEATURE SELECTION METHODS. Technical report, 2022. URL www.jatit.org.
- [65] Suvajit Dutta, Bonthala C.S. Manideep, Syed Muzamil Basha, Ronnie D. Caytiles, and N. Ch S.N. Iyengar. Classification of diabetic retinopathy images by using deep learning models. *International Journal of Grid and Distributed Computing*, 11(1):89–106, 2018. ISSN 22076379. doi: 10.14257/ijgdc.2018.11.1.09.
- [66] Shahnawaz Ayoub, Mohiuddin Ali Khan, Vaishali Prashant Jadhav, Harishchander Anandaram, T. Ch Anil Kumar, Faheem Ahmad Reegu, Deepak Motwani, Ashok Kumar Shrivastava, and Roviell Berhane. Minimized Computations of Deep Learning Technique for Early Diagnosis of Diabetic Retinopathy Using IoT-Based Medical Devices. *Computational intelligence and neuroscience*, 2022:7040141, 2022. ISSN 16875273. doi: 10.1155/2022/7040141.
- [67] Kwang Sun Ryu, Sang Won Lee, Erdenebileg Batbaatar, Jae Wook Lee, Kui Son Choi, and Hyo Soung Cha. A deep learning model for estimation of patients

- with undiagnosed diabetes. *Applied Sciences (Switzerland)*, 10(1), jan 2020. ISSN 20763417. doi: 10.3390/app10010421.
- [68] Bharat Tayappa Jadhav, K P Mali, B T Jadhav, and I K Mujawar. Study of Diabetic Retinopathy Detection Using Deep Learning Techniques Delopment of expert system for diabetes patients View project Study of Diabetic Retinopathy Detection Using Deep Learning Techniques. Technical report. URL <https://www.researchgate.net/publication/359760459>.
- [69] María Teresa García-Ordás, Carmen Benavides, José Alberto Benítez-Andrades, Héctor Alaiz-Moretón, and Isaías García-Rodríguez. Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Computer Methods and Programs in Biomedicine*, 202(February 2021), 2021. ISSN 18727565. doi: 10.1016/j.cmpb.2021.105968.
- [70] K. Kannadasan, Damodar Reddy Edla, and Venkatanareshbabu Kuppili. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clinical Epidemiology and Global Health*, 7(4):530–535, 2019. ISSN 22133984. doi: 10.1016/j.cegh.2018.12.004.
- [71] Soham Mehta. Diabetes Prediction using Deep Neural Network Diabetes Prediction Using Deep Neural Network View project DIABETES PREDICTION USING DEEP NEURAL NETWORK. Technical report. URL www.irjmetts.com.
- [72] Bala Manoj Kumar P, Srinivasa Perumal R, Nadesh R K, and Arivuselvan K. Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier. *International Journal of Cognitive Computing in Engineering*, 1:55–61, jun 2020. ISSN 26663074. doi: 10.1016/j.ijcce.2020.10.002.
- [73] Toshita Sharma and Manan Shah. A comprehensive review of machine learning techniques on diabetes detection. 2021.
- [74] Mathematics Education, A Prakash, O Vignesh, R Suneetha Rani, and S Abinayaa. An Ensemble Technique for Early Prediction of Type 2 Diabetes Mellitus – A Normalization Approach. 12(9):2136–2143, 2021.
- [75] Muhammad Nabeel, Shumaila Majeed, Mazhar Javed Awan, Hooria Muslih-Uddin, Mashal Wasique, and Rabia Nasir. Review on effective disease prediction through data mining techniques. *International Journal on Electrical Engineering and Informatics*, 13(3):717–733, 2021. ISSN 20875886. doi: 10.15676/IJEEI.2021.13.3.13.

- [76] Avraham Adler. Using Machine Learning Techniques to Identify Key Risk Factors for Diabetes and Undiagnosed Diabetes. 2021.
- [77] Farrukh Aslam Khan, Senior Member, Khan Zeb, Mabrook Al-rakhami, and Abdelouahid Derhab. Detection and Prediction of Diabetes Using Data Mining : A Comprehensive Review. pages 43711–43735, 2021. doi: 10.1109/ACCESS.2021.3059343.
- [78] Sujithra Sankar and S Sathyalakshmi. PREDICTION OF ENDOCRINE DISORDERS USING MACHINE LEARNING CLASSIFICATION ALGORITHMS : A COMPREHENSIVE. (July 2021):151–163, . doi: 10.17605/OSF.IO/FYS93.
- [79] A Comparative Study. applied sciences Data Mining Techniques for Early Diagnosis of Diabetes :. pages 1–12, 2021.
- [80] Fikirte Girma Woldemichael and Sumitra Menaria. Prediction of Diabetes Using Data Mining Techniques. *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, (Icoei):414–418, 2018. doi: 10.1109/ICOEI.2018.8553959.
- [81] January February, Oluwafemi Samuel Abe, Olumide O Obe, Olutayo K Boyinbode, and Olagbuji N Biodun. Classifier Algorithms and Ensemble Models for Diabetes Mellitus Prediction: A Review. *International Journal of Advanced Trends in Computer Science and Engineering*, 10(1):430–439, 2021. doi: 10.30534/ijatcse/2021/641012021.
- [82] Tetiana Dudkina, Ievgen Meniailov, Kseniia Bazilevych, and Serhii Krivtsov. Classification and Prediction of Diabetes Disease using Decision Tree Method. 2836: 0–1, 2021.
- [83] M R ASengamuthu, M R Birami, and ... Various Data Mining Techniques Analysis to Predict Diabetes Mellitus. *Int Res J Eng Technol ...*, pages 676–679, 2018. URL <https://www.academia.edu/download/57013446/IRJET-V5I5134.pdf>.
- [84] Jafar Abdollahi and Babak Nouri-moghaddam. Hybrid stacked ensemble combined with genetic algorithms for Prediction of Diabetes.
- [85] V Prudhvi. PREDICTION OF DIABETES MELLITUS USING RBF NEURAL MODEL AND GENETIC ALGORITHM. 32(3):1524–1531, 2021.
- [86] Saloni Kumari, Deepika Kumar, and Mamta Mittal. International Journal of Cognitive Computing in Engineering An ensemble approach for classification and

- prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, 2(November 2020):40–46, 2021. ISSN 2666-3074. doi: 10.1016/j.ijcce.2021.01.001. URL <https://doi.org/10.1016/j.ijcce.2021.01.001>.
- [87] Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz, Ayaz Hussain, Muhammad Awais, and Muhammad Mehdi. Informatics in Medicine Unlocked A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16(January):100204, 2019. ISSN 2352-9148. doi: 10.1016/j.imu.2019.100204. URL <https://doi.org/10.1016/j.imu.2019.100204>.
- [88] Amani Yahyaoui. A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. (2):1–4, 2019.
- [89] N Sneha and Tarun Gangil. Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0175-6. URL <https://doi.org/10.1186/s40537-019-0175-6>.
- [90] Gopi Battineni, Getu Gamo Sagaro, Chintalapudi Nalini, Francesco Amenta, and Seyed Khosrow Tayebati. Comparative Machine-Learning Approach : A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. pages 1–11, 2019.
- [91] Amith Khandakar, Muhammad E H Chowdhury, Mamun Bin, Ibne Reaz, Sawal Hamid, Anwarul Hasan, Serkan Kiranyaz, Tawsifur Rahman, Rashad Alfkey, Ahmad Ashrif, A Bakar, and Rayaz A Malik. A machine learning model for early detection of diabetic foot using thermogram images. *Computers in Biology and Medicine*, 137(September):104838, 2021. ISSN 0010-4825. doi: 10.1016/j.compbimed.2021.104838. URL <https://doi.org/10.1016/j.compbimed.2021.104838>.
- [92] Deepti Sisodia and Dilip Singh Sisodia. ScienceDirect Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, 132(Iccids):1578–1585, 2018. ISSN 1877-0509. doi: 10.1016/j.procs.2018.05.122. URL <https://doi.org/10.1016/j.procs.2018.05.122>.
- [93] Neha Prerna Tigga and Shruti Garg. ScienceDirect ScienceDirect Prediction of Type 2 Diabetes using Machine Learning Prediction of Type 2 Diabetes using Machine Learning Classification Methods Classification Methods. *Procedia Computer Science*, 167(2019):706–716, 2020. ISSN 1877-0509. doi: 10.1016/j.procs.2020.03.336. URL <https://doi.org/10.1016/j.procs.2020.03.336>.

- [94] Dola Das, Eklas Hossain, and Mahmudul Hasan. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. 8, 2020. doi: 10.1109/ACCESS.2020.2989857.
- [95] Minakhi Rout, Jitendra Kumar Rout, and Himansu Das. *Nature Inspired Computing for Data Science*. 2020. ISBN 9783030338190.
- [96] Harleen Kaur and Vinita Kumari. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, 18(1-2): 90–100, 2022. ISSN 22108327. doi: 10.1016/j.aci.2018.12.004.
- [97] Oladosu Oyebisi Oladimeji, Abimbola Oladimeji, and Olayanju Oladimeji. Classification models for likelihood prediction of diabetes at early stage using feature selection. *Applied Computing and Informatics*, 20(3-4):279–286, 2021. ISSN 22108327. doi: 10.1108/ACI-01-2021-0022.
- [98] Ibrahim Mahmood Ibrahim and Adnan Mohsin Abdulazeez. The Role of Machine Learning Algorithms for Diagnosing Diseases. 02(01):10–19, 2021. doi: 10.38094/jastt20179.
- [99] Minyechil Alehegn. Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm. 118(9):871–878, 2018.
- [100] Jingyuan Wang, Xiujuan Chen, and Kai Chen. Machine Learning Approaches for Early Prediction of Gestational Diabetes Mellitus Based on Prospective Cohort Study. pages 1–14, 2021.
- [101] Aditya Saxena, Megha Jain, and Prashant Shrivastava. Data Mining Techniques Based Diabetes Prediction. 7626(2):29–35, 2021. doi: 10.35940/ijainn.B1012.041221.
- [102] R Murugadoss. Early Prediction of Diabetes Using Deep Learning Convolution Neural Network and Harris Hawks Optimization. 1:88–100, 2021.
- [103] F. M. Javed Mehedi Shamrat, Md Abu Raihan, A. K.M.Sazzadur Rahman, Imran Mahmud, and Rozina Akter. An analysis on breast disease prediction using machine learning approaches. *International Journal of Scientific and Technology Research*, 9(2):2450–2455, 2020. ISSN 22778616.
- [104] Mohamed Chetoui, Moulay A Akhloufi, and Mustapha Kardouchi. Diabetic Retinopathy Detection Using Machine Learning and Texture Features. 2018.

- [105] Ahmed J Aljaaf, Dhiya Al-jumeily, Hussein M Haglan, Mohamed Alloghani, Thar Baker, and Abir J Hussain. Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics. *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–9, 2018.
- [106] Sajratul Yakin Rubaiat, Monibor Rahman, and Kamrul Hasan. Important Feature Selection Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection. (1):1–6, 2018.
- [107] V Krishnapraseeda. Predictive Analytics on Diabetes Data using Machine Learning Techniques. pages 1670–1673, 2021.
- [108] Rahatara Ferdousi, M Anwar Hossain, and Abdulmotaleb El Saddik. Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS. *IEEE Access*, 9:96823–96837, 2021. doi: 10.1109/ACCESS.2021.3094063.
- [109] Faisal Faruque. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–4, 2019.
- [110] Janhavi R Raut. PERFORMANCE EVALUATION OF VARIOUS SUPERVISED MACHINE LEARNING ALGORITHMS FOR DIABETES. 7(8):4921–4925, 2020.
- [111] P Moksha Sri Sai and G Anuradha. Machine Learning. (Iccmc):770–775, 2020. doi: 10.1109/ICCMC48092.2020.ICCMC-000143.
- [112] Ayman Mir and Sudhir N Dhage. Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6, 2018.
- [113] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, and Munam Ali Shah. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. *2018 24th International Conference on Automation and Computing (ICAC)*, (September):1–6, 2018. doi: 10.23919/ICOnAC.2018.8748992.
- [114] K Vijiyakumar, B Lavanya, I Nirmala, and S Sofia Caroline. Random Forest Algorithm for the Prediction of Diabetes. *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pages 1–5, 2019.

- [115] Satoru Tanioka, Fujimaro Ishida, Fumi Nakano, Fumihiro Kawakita, and Hideki Kanamaru. Machine Learning Analysis of Matricellular Proteins and Clinical Variables for Early Prediction of Delayed Cerebral Ischemia After Aneurysmal Subarachnoid Hemorrhage. (MI):9–12, 2019.
- [116] Leon Kopitar, Primož Kocbek, Leona Cilar, Aziz Sheikh, and Gregor Stiglic. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, 10(1), dec 2020. ISSN 20452322. doi: 10.1038/s41598-020-68771-z.
- [117] J Omana and M Moorthi. Prediction Of Diabetes Mellitus Using Measure Of Insulin Resistance : A Combined Classifier Approach. 12(11):4793–4801, 2021.
- [118] Samarjeet Borah. *Soft Computing Techniques and Applications*. Number Ic3. 2020. ISBN 9789811573934.
- [119] Qingqing Xu, Liye Wang, and Sujit S Sansgiry. A systematic literature review of predicting diabetic retinopathy , nephropathy and neuropathy in patients with type 1 diabetes using machine learning. (MI), 2019. doi: 10.21037/jmai.2019.10.04.
- [120] Omar AlShorman, Buthaynah AlShorman, and Fahed Alkahtani. A review of wearable sensors based monitoring with daily physical activity to manage type 2 diabetes. *International Journal of Electrical and Computer Engineering*, 11(1): 646–653, 2021. ISSN 20888708. doi: 10.11591/ijece.v11i1.pp646-653.
- [121] Pronab Ghosh, Sami Azam, Asif Karim, Mehedi Hassan, Kuber Roy, and Mirjam Jonkman. A comparative study of different machine learning tools in detecting diabetes. *Procedia Computer Science*, 192:467–477, 2021. ISSN 18770509. doi: 10.1016/j.procs.2021.08.048. URL <https://doi.org/10.1016/j.procs.2021.08.048>.
- [122] P. Bharath Kumar Chowdary and R. Udaya Kumar. An Effective Approach for Detecting Diabetes using Deep Learning Techniques based on Convolutional LSTM Networks. *International Journal of Advanced Computer Science and Applications*, 12(4):519–525, 2021. ISSN 21565570. doi: 10.14569/IJACSA.2021.0120466.
- [123] Harshil Thakkar, Vaishnavi Shah, Hiteshri Yagnik, and Manan Shah. Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. *Clinical eHealth*, 4(2021):12–23, 2021. ISSN 25889141. doi: 10.1016/j.ceh.2020.11.001. URL <https://doi.org/10.1016/j.ceh.2020.11.001>.

- [124] Ovass Shafi Zargar, Avinash Baghat, and Tawseef Ahmed Teli. A DNN Model for Diabetes Mellitus Prediction on PIMA Dataset. (2):1–9, 2022. URL <https://infocomp.dcc.ufla.br/index.php/infocomp/article/view/2476>.
- [125] Huma Naz and Sachin Ahuja. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes and Metabolic Disorders*, 19(1):391–403, jun 2020. ISSN 22516581. doi: 10.1007/s40200-020-00520-5.
- [126] K. Thaiyalnayaki. Classification of diabetes using deep learning and svm techniques. *International Journal of Current Research and Review*, 13(1):146–149, jan 2021. ISSN 09755241. doi: 10.31782/IJCRR.2021.13127.
- [127] Safial Islam Ayon and Md. Milon Islam. Diabetes Prediction: A Deep Learning Approach. *International Journal of Information Engineering and Electronic Business*, 11(2):21–27, mar 2019. ISSN 20749023. doi: 10.5815/ijieeb.2019.02.03. URL <http://www.mecs-press.org/ijieeb/ijieeb-v11-n2/v11n2-3.html>.
- [128] Pélagie Houngué and Annie Ghylaine Bigirimana. Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network. *Journal of Computer and Communications*, 10(11):15–28, 2022. ISSN 2327-5219. doi: 10.4236/jcc.2022.1011002.
- [129] Sujithra Sankar and S Sathyalakshmi. PREDICTION OF ENDOCRINE DISORDERS USING MACHINE LEARNING CLASSIFICATION ALGORITHMS : A COMPREHENSIVE. (July 2021):151–163, . doi: 10.17605/OSF.IO/FYS93.
- [130] January February, Oluwafemi Samuel Abe, Olumide O Obe, Olutayo K Boyinbode, and Olagbuji N Biodun. Classifier Algorithms and Ensemble Models for Diabetes Mellitus Prediction: A Review. *International Journal of Advanced Trends in Computer Science and Engineering*, 10(1):430–439, 2021. doi: 10.30534/ijatcse/2021/641012021.
- [131] Oladosu Oyebisi Oladimeji, Abimbola Oladimeji, and Olayanju Oladimeji. Classification models for likelihood prediction of diabetes at early stage using feature selection. *Applied Computing and Informatics*, 20(3-4):279–286, 2024. ISSN 22108327. doi: 10.1108/ACI-01-2021-0022.
- [132] Jingyuan Wang, Xiujuan Chen, and Kai Chen. Machine Learning Approaches for Early Prediction of Gestational Diabetes Mellitus Based on Prospective Cohort Study. pages 1–14.

- [133] Hafiz Farooq Ahmad, Hamid Mukhtar, Hesham Alaqail, Mohamed Seliaman, and Abdulaziz Alhumam. Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Applied Sciences (Switzerland)*, 11(3):1–18, feb 2021. ISSN 20763417. doi: 10.3390/app11031173.
- [134] Thippa Reddy, Gadekallu Neelu, Khare Sweta, and Bhattacharya Saurabh. Deep neural networks to predict diabetic retinopathy. *Journal of Ambient Intelligence and Humanized Computing*, (0123456789), 2020. ISSN 1868-5145. doi: 10.1007/s12652-020-01963-7. URL <https://doi.org/10.1007/s12652-020-01963-7>.
- [135] T P Latchoumi, J Dayanika, and G Archana. A Comparative Study of Machine Learning Algorithms using Quick-Witted Diabetic Prevention. Technical report, 2021. URL <http://annalsofrscb.ro>.
- [136] Borys Tymchenko, Philip Marchenko, and Dmitry Spodarets. Deep Learning Approach to Diabetic Retinopathy Detection. mar 2020. URL <http://arxiv.org/abs/2003.02261>.
- [137] Thippa Reddy Gadekallu, Neelu Khare, Sweta Bhattacharya, Saurabh Singh, Praveen Kumar Reddy Maddikunta, and Gautam Srivastava. Deep neural networks to predict diabetic retinopathy. *Journal of Ambient Intelligence and Humanized Computing*, 2020. ISSN 18685145. doi: 10.1007/s12652-020-01963-7.
- [138] Motiur Rahman, Dilshad Islam, Rokeya Jahan Mukti, and Indrajit Saha. A deep learning approach based on convolutional LSTM for detecting diabetes. *Computational Biology and Chemistry*, 88, oct 2020. ISSN 14769271. doi: 10.1016/j.compbiolchem.2020.107329.
- [139] Akm Ashiquzzaman, Abdul Kawsar Tushar, Md Rashedul Islam, Dongkoo Shon, Kichang Im, Jeong Ho Park, Dong Sun Lim, and Jongmyon Kim. Reduction of overfitting in diabetes prediction using deep learning neural network. In *Lecture Notes in Electrical Engineering*, volume 449, pages 35–43. Springer Verlag, 2017. ISBN 9789811064500. doi: 10.1007/978-981-10-6451-7_5.
- [140] Tatsuhiko Naito, Ken Suzuki, Jun Hirata, Yoichiro Kamatani, Koichi Matsuda, Tatsushi Toda, and Yukinori Okada. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nature Communications*, 12(1), dec 2021. ISSN 20411723. doi: 10.1038/s41467-021-21975-x.
- [141] Emet D. Schneiderman, Charles J. Kowalski, and Stephen M. Willis. Regression imputation of missing values in longitudinal data sets. *International*

- Journal of Bio-Medical Computing*, 32(2):121–133, 1993. ISSN 00207101. doi: 10.1016/0020-7101(93)90051-7.
- [142] Wei Chao Lin and Chih Fong Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2):1487–1509, 2020. ISSN 15737462. doi: 10.1007/s10462-019-09709-4. URL <https://doi.org/10.1007/s10462-019-09709-4>.
- [143] Ovass Shafi, Jahangir S Sidiq, Tawseef Ahmed Teli, and Kashmir . EFFECT OF PRE-PROCESSING TECHNIQUES IN PREDICTING DIABETES MELLITUS WITH FOCUS ON ARTIFICIAL NEURAL NETWORK. Technical Report 8, 2022.
- [144] Munish Khanna, Law Kumar Singh, Shankar Thawkar, and Mayur Goyal. Deep learning based computer-aided automatic prediction and grading system for diabetic retinopathy. *Multimedia Tools and Applications*, mar 2023. ISSN 1380-7501. doi: 10.1007/s11042-023-14970-5. URL <https://link.springer.com/10.1007/s11042-023-14970-5>.
- [145] Taiyu Zhu, Kezhi Li, Pau Herrero, and Pantelis Georgiou. Deep Learning for Diabetes: A Systematic Review, jul 2021. ISSN 21682208.
- [146] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022. ISSN 2666285X. doi: 10.1016/j.gltp.2022.04.020.
- [147] R. M. Anjana, R. Pradeepa, and Deepa. Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: Phase i results of the Indian Council of Medical Research-INDia DIABetes (ICMR-INDIAB) study. *Diabetologia*, 54(12):3022–3027, 2011. ISSN 0012186X. doi: 10.1007/s00125-011-2291-5.
- [148] Ranjit Mohan Anjana and Deepa. Prevalence of diabetes and prediabetes in 15 states of India: results from the ICMR–INDIAB population-based cross-sectional study. *The Lancet Diabetes and Endocrinology*, 5(8):585–596, 2017. ISSN 22138595. doi: 10.1016/S2213-8587(17)30174-2.
- [149] Ashish Bora and Balasubramanian. Predicting the risk of developing diabetic retinopathy using deep learning. *The Lancet Digital Health*, 3(1):e10–e19, jan 2021. ISSN 25897500. doi: 10.1016/S2589-7500(20)30250-8.

- [150] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97(xxxx): 105524, 2020. ISSN 15684946. doi: 10.1016/j.asoc.2019.105524. URL <https://doi.org/10.1016/j.asoc.2019.105524>.
- [151] Ovass Shafi Zargar, Avinash Bhagat, and Tawseef Ahmed Teli. Feature Selection, Importance and Missing Value Imputation in Diabetes Mellitus Prediction. *Proceedings of the 17th INDIACom; 2023 10th International Conference on Computing for Sustainable Global Development, INDIACom 2023*, pages 914–919, 2023.
- [152] Vipin Kumar. Feature Selection: A literature Review. *The Smart Computing Review*, 4(3), 2014. doi: 10.6029/smarter.2014.03.007.
- [153] Ovass Shafi Zargar Avinash Bhagat, Tawseef Ahmed Teli. A Deep Learning-Based Diabetes Diagnosis Model on PIMA Image Dataset. *Journal of Electrical Systems*, 20(3s):1276–1289, 2024. doi: 10.52783/jes.1444.
- [154] Suyash Srivastava, Lokesh Sharma, Vijeta Sharma, Ajai Kumar, and Hemant Darbari. *Prediction of Diabetes Using Artificial Neural Network Approach: ICo-EVCI 2018, India*, pages 679–687. 01 2019. ISBN 978-981-13-1641-8. doi: 10.1007/978-981-13-1642-5_59.
- [155] G. Swapna, R. Vinayakumar, and K. P. Soman. Diabetes detection using deep learning algorithms. *ICT Express*, 4(4):243–246, 2018. ISSN 24059595. doi: 10.1016/j.icte.2018.10.005. URL <https://doi.org/10.1016/j.icte.2018.10.005>.