OPTIMISATION OF DEEP LEARNING MODELS USING FLOCK ALGORITHM IN IMPROVING FEATURE MODELLING AND DIABETES CLASSIFICATION

By

BALASUBRAMANIYAN DIVAGER

Thesis submitted to the School of Graduate Studies, University Putra Malaysia, in fulfilment of the Requirements for the Degree of Doctor of Philosophy

June 2024

FSKTM 2024 19

Abstract of thesis presented to the Senate of University Putra Malaysia in fulfilment of the requirement for the Degree of Doctor of Philosophy

OPTIMISATION OF DEEP LEARNING MODELS USING FLOCK ALGORITHM IN IMPROVING FEATURE MODELLING AND DIABETES CLASSIFICATION

By

BALASUBRAMANIYAN DIVAGER

June 2024

Chairman : Nor Azura binti Husin, PhD
Faculty : Computer Science and Information Technology

Diabetic illness is a widespread health issue that affects people of both genders. The diabetic disease creates several diseases, such as heart problems, stroke, etc. Therefore, patients are doing a self-assessment process to identify the diabetic disease in the earlier stage. However, the self-assessment process fails to detect most cases, leading to increased complexity. The prediction model requires robust training model due to the insufficient data assimilation that leads to the generalization problem and minimize the prediction issues. Another important problem is imbalanced data which occurred due to the uneven distribution of diabetic classes (non-diabetic cases and diabetic case). The imbalance problem having influence on the learning ability and accuracy; Therefore, an automatic detection system is created to identify diabetic disease at an earlier stage. However, the existing diabetic prediction models fail to address the data assimilation, imbalance data, information loss, and accuracy issues. Therefore, this research proposed a method called the flock-optimized assimilated deep learning model (FOADLM) to improve diabetic detection efficiency. The flock optimization

approach maximizes model convergences and minimize training time which leads to improve the diabetic detection efficiency. The PIMA Indian Dataset, Azure Dataset, and OhioT1DM Dataset information are used to evaluate the accuracy of FOADLM. Initially, the Gaussian Filtering is used to process the data by using the kernel value to filter out irrelevant information. Then neural convolution model is integrated with the flock optimization technique to determine the features. The effective utilization of optimization techniques recognizes the relationship between the data, minimize the information loss, and improve the overall data accuracy. Then data sampling techniques is applied by combine with the flock optimization in Convolution Neural Networks (CNN)'s first layer achieves the system flexibility. During the analysis, pre-dominant and latent features are utilized in the classification that minimize the computation complexity and improves the overall recognition accuracy. The Long Short Term Memory Neural Networks (LSTM) network analyses the discovered features to identify diabetes effectively. Thus, the system achieves 99.23% accuracy and minimum error rate (0.1) as findings, surpassing the previous works and various diabetic detection systems.

Keyword: Diabetic Disease, Data Assimilation, Data Sampling, Imbalance Issues, Long-Short Term Memory Neural Networks.

SDG: GOAL 3: Good Health and Well-Being, GOAL 9: Industry, Innovation, and Infrastructure, GOAL 10: Reduced Inequalities

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia Sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENGOPTIMUMAN MODEL PEMBELAJARAN MENDALAM MENGGUNAKAN KAWANAN ALGORITMA DALAM MENINGKATKAN CIRI-CIRI PEMODELAN DAN KLASIFIKASI DIABETES

Oleh

BALASUBRAMANIYAN DIVAGER

June 2024

Pengerusi : Nor Azura binti Husin, PhD
Fakulti : Sains Komputer dan Teknologi Maklumat

Penyakit diabetes adalah isu kesihatan yang meluas yang memberi kesan kepada manuasia kepada kedua-dua jantina. Penyakit diabetes menimbulkan beberapa penyakit, seperti masalah jantung, strok dan sebagainya. Oleh itu, pesakit perlu melakukan proses penilaian kendiri untuk mengenal pasti penyakit diabetes pada peringkat awal. Walau bagaimanapun, proses penilaian kendiri gagal mengesan kebanyakan kes, yang membawa kepada peningkatan yang merumitkan. Oleh itu, sistem pengesanan automatik diwujudkan untuk mengenal pasti penyakit diabetes pada peringkat awal. Walau bagaimanapun, sistem ramalan diabetes sedia ada gagal menangani asimilasi data, ketidakseimbangan data, kehilangan maklumat dan isu ketepatan. Oleh itu, penyelidikan ini mencadangkan satu kaedah yang dipanggil model pembelajaran mendalam berasimilasi yang dioptimumkan oleh kumpulan (FOADLM) untuk meningkatkan kecekapan pengesanan diabetes. Pendekatan pengoptimuman kumpulan memaksimumkan penumpuan model dan meminimumkan masa latihan yang membawa kepada meningkatkan kecekapan pengesanan diabetes. Maklumat

PIMA Indian, Azure dan OhioT1DM Dataset digunakan untuk menilai ketepatan FOADLM. Penapisan Gaussian digunakan untuk memproses data yang diperolehi dengan menggunakan nilai kernel untuk menapis maklumat yang tidak berkaitan. Kemudian model lilitan saraf disepadukan dengan teknik pengoptimuman kawanan untuk menentukan ciri. Penggunaan teknik pengoptimuman yang berkesan mengiktiraf hubungan antara data, meminimumkan kehilangan maklumat, dan meningkatkan ketepatan data keseluruhan. Rangkaian LSTM menggunakan ciri yang ditemui untuk mengenal pasti diabetes dengan ketepatan sehingga 99.23%, mengatasi kerja terdahulu dan pelbagai sistem pengesanan diabetes.

Kata kunci: Asimilasi Data, Isu Ketidakseimbangan, Pensampelan Data, Penyakit Diabetik, Rangkaian Saraf Ingatan Jangka Pendek.

SDG: MATLAMAT 3: Kesihatan dan Kesejahteraan yang Baik, MATLAMAT 9: Industri, Inovasi dan Infrastruktur, MATLAMAT 10: Ketaksamaan Berkurangan

ACKNOWLEDGEMENTS

Firstly, I would like to thank the almighty and my mother for the good health, wisdom, opportunities and prosperity. My sincere thanks to my mother, she sacrificed her all life to sculpt me and challenged the society to create this wonderful career for me as a single parent and also I would like to thank my all family members for the support.

Secondly, I would like to thank my supervisor, Dr. Nor Azura Husin, for guiding me from the beginning and always supported, motivated me. I would also like to thank my co-supervisors Assoc. Professor Dr. Norwati Mustafa, Assoc. Professor Dr. Nurfadhlina Mohd sharef and Assoc. Professor Dr. Teh Noranis Mohd Aris for contributing time and supported with valuable comments during this period of my journey.

Lastly I would like to express my thanks and gratitude to my brother Dr Shakeel for the motivation and support in all my challenges. Many more thanks to all the well-wishers really cared, motivated and appreciated for my success.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Nor Azura binti Husin, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Norwati binti Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Nurfadhlina binti Mohd Sharef, PhD
Professor Ts.
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Teh Noranis binti Mohd Aris, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 24 February 2025

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC	Ant Bee Colony Optimization
AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
CGM	Continuous Glucose Monitoring
CNN	Convolution Neural Networks
DCNN	Deep Convolutional Neural System
DNN	Deep Neural Networks
ELA	Ensemble Learning Approach
FOADLM	Flock Optimization Assimilated Deep Learning Model
GA	Genetic Algorithm
GODNL	Genetically Optimized Deep Neural Learning
IGDBN	Iterative Golden Section Optimized Deep Belief Neural Network
IPSO-LSTM	Improved Particle Swarm Optimization Optimizing Long Short-Term Memory Network
LDA	Linear Discriminate Analysis
LOF	Local Outlier Factor
LSTM	Long Short Term Memory Networks
MCMTTP	Multi-Centre and Multi-Threshold-Based Ternary Pattern-based Voice Algorithm
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RNN	Recurrent Neural Networks
SMBG	Self-Observing Blood Glucose
TOSMN	Tubu Optimized Sequence Modular Network
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

1.1 Background

Diabetic disease (Balwan et al., 2021) is one of the serious diseases that is the reason for several diseases such as heart attack, chronic respiratory, and cardiovascular disease. Diabetic disease requires the initial diagnosis and treatment procedure to recover the patient from the critical situation. Several researchers use machine learning and data mining techniques (Islam et al., 2020) to predict diabetic disease detection. The issue under consideration machine learning methods to forecast the probability of an individual's susceptibility to diabetes. Diabetes, an ongoing condition defined by hyperglycaemia, necessitates timely identification to facilitate efficient control and mitigate the risk of associated complications. The task at hand involves the acquisition of an extensive dataset that encompasses relevant characteristics such as personal qualities, physiological signs, and levels of biomarkers. Following the first data pre-treatment steps, which include cleaning, normalization, and feature engineering, the subsequent phase involves the selection of appropriate machine learning models. The models, encompassing logistic regression and neural networks, will undergo training on a partitioned dataset and be assessed using criteria such as accuracy and precision. The machine learning model uses the metaheuristic models for improving the diabetic predictions system performance. The diabetic model helps to manage the non-convex spaces, complex and high-dimensional data. Several meta-heuristic optimization procedures such as genetic algorithm, particle swarm optimization, harmony search,

ant bee colony and bacterial foraging optimization etc. are utilized for improving the diabetic prediction model performance. Diabetic identification model handles high dimensional data with high parameter and settings. The model has been designed with the help of meta-heuristic optimization method that handle the large search space and navigate the complex problems. In addition, the optimization model detects the optimal configurations which helps to maximize the overall prediction rate. The prediction model facing the non-convex optimization issue which means, system having the multiple local optimal solution. The non-convex optimization in diabetic classification is objective function consists of several maxima and minima which leads to create the difficulties while selecting the global solution. Therefore, the non-convex optimization process requires the optimization techniques to ensure the reliability and accuracy during the prediction. Therefore, the system gets disrupted while selecting the suboptimal solution which is optimized using the optimization methods. In addition, optimization methods address the convergence issues during the data handling and training process. The optimization model balances the deviation between the classes that reduces the data imbalance and insufficient data assimilation issues effectively. The meta-heuristic optimization model based created diabetic prediction model ensure the robustness while processing the diverse data. The primary goal is to implement a precise and understandable prediction model in practical healthcare systems, facilitating the prompt detection of individuals who are susceptible to risks and facilitating well-informed actions. The continuous Improvement of this model will guarantee its sustained effectiveness in facilitating the early detection of diabetes and providing individualized patient treatment.

1.2 Problem Statement

In diabetic prediction model, insufficient data assimilation is defined as the inadequacy representation of data in the model training phase. The insufficient data assimilation occurs during the data does not covers the entire diabetic cases. The data assimilation issues create the generalization problem and minimize the diabetic detection accuracy while analysing new data. Another important problem is imbalanced data which occurred due to the uneven distribution of diabetic classes (non-diabetic cases and diabetic case). The imbalance problem having influence on the learning ability and accuracy; therefore, the model needs to understand the entire patterns while exploring diabetic data. Li K. et al. (2019) investigated the application of a Convolutional Recurrent Neural Network for predicting glucose levels. Convolutional layers were incorporated to reduce the number of parameters efficiently. The system lacks sufficient flexibility to effectively categorize parameters due to inadequate data sampling and optimization procedures. The pooling layer used in the model loses valuable information and ignores to examine the relationship between each feature that seriously affects the disease recognition efficiency. The current model does not address the issues of convergence, non-converging optimization, complicated predictions, and latent and predominant (complex patterns and significant features) feature extraction. The convergence problem creates the difficulties while selecting the optimal and stable solution during the training phase of learning model. The convergence issues cause the system facing struggle to attains the further iteration which means system does not have any significant improvement in performance. The convergence problem is the main reason for higher training time and failures while selecting optimal solutions. Due to the low-depth and complexity of the current model. This could be missing out

some latent features which will appreciate those hidden patterns in data. The latent feature has an effect on performance by reducing accuracy to smaller problems and decreasing generalization, especially when the relationships between data is more complex, or patterns are less obvious. However, each model has shortcomings in that might struggle to capture the hidden features of the status independently due to its constraints and limitations such as overfitting new data will lead into poor generalization when encountering unseen data or not accurately displaying latent patterns which result into misclassification if considering non-linear relationship between variables. This limitation can cause overfitting as well where the model is good at training data but could miss out on meaningful underlying trends required for appropriate real-world predictions. Therefore, the diabetic system requires the effective ad optimized model to generalize the solution because this helps to ensure the system reliability and robustness.

Based on the above-discussed problems, the existing system failed to produce accuracy due to insufficient data assimilation and imbalanced data. Data assimilation is the way of integrating the practical data into the computation model for improving the prediction accuracy. The effective integration of this process updates the model for getting the reliable outputs. Data imbalance is occurred during the skewed distribution which affects the classifier performance.

The existing modified Long Short-Term Memory (Hou, L., et al. 2019) facing difficulties while initialize the random weight initialization, which is prone to data overfitting problems. Since dropout is often used as a regularization strategy, Long Short-Term Memory units' input and periodic connections are often removed from

activation and weights upgrades during training. He, J et al. (2020) pay attention to optimized techniques and kernel canonical correlation analysis to predict the blood glucose level. The correlation between real-time and predicted glucose values shows up deviations. These deviations can be reduced by assimilating practical and effective optimization techniques for disease classification. Wang et al. (2020) introduced improved and optimized techniques for predicting blood glucose. However, due to its incorporation of hard-level efficiency and learning approaches, the suggested solution has a high execution complexity. Li K. et al. (2019) pay attention to Convolutional Recurrent Neural Networks for Glucose Prediction. The modifications are common regardless of the data stream and quantity. This results in non-adapting computations, requiring multiple repeated validations (high prediction horizons), and data overfitting problem also occurs. Li, K. et al. (2020) developed a personalized deep neural network with a probabilistic distribution for glucose prediction. However, the model required large amounts of high-quality training data to maintain accuracy over prolonged periods. Additionally, the increased validation instances led to longer time lags and high validation times.

1.3 Research Questions

1. How the data assimilation process to assure the maximum diabetic classification accuracy and how it creates the impact on accuracy and generalization?
2. What techniques are applied to address the data imbalance and overfitting issues while performing the new diabetic cases to attains the maximum accuracy?
3. How can the challenges of model convergence, overfitting, and effective feature extraction be mitigated in glucose prediction models, particularly in handling noise, missing data, and irregular sampling for optimal performance?

1.4 Objective of the Study

The proposed research aim is to develop a diabetic disease prediction model based on Flock Optimization Assimilated Deep Learning Model (FOADLM) to improve the diabetic prediction accuracy.

1. To assess system flexibility by applying data sampling and Flock optimization in the first layer of a CNN, aiming to improve the classification accuracy and flexibility by analysing the relationships between data features.
2. To design the diabetic prediction model by investigating the latent and pre-dominant features using the sampling and flock optimization techniques. Successfully extracting features diminishes the deviation between real-time and predicted glucose values.
3. To develop the flock-optimized neural network to provide the best solution while handling the imbalanced and assimilating data. This process helps to minimize the computation complexity and time complexity issues.

1.5 Scope of the Study

The research aims to address the difficulties and restrictions linked with the diabetic classification process. First the work concentrates on the assimilation techniques to handling the data assimilation issues to create the benchmark analysis for future techniques. Then, less-variation prediction techniques are explored that works along with the user-specific information and multi-feature association. The variation analysis maximizes the prediction accuracy. The hard-level computations will be suppressed irrespective of the different data and patient input features. This means the assimilation must be adaptable, aiding the overall performance enhancement. The defined model uses the training and validation model, which helps to train the numerous data features. The training process helps to find new patterns from the learned features, effectively minimizing the complexity. Proposing a synchronization identifying optimization

regardless of the attribute and detecting its impact over the assimilated proportion. This would help adjust the proportion of the techniques in retaining the prediction accuracy. Here, the application of flock metaheuristic optimization is important since the application optimize elements of feature selection in high dimensional complex datasets for overfitting minimization as well as generalization improvement. As in biological flocks, this optimization technique becomes dynamic and readily handles the varied and continuous nature of patient data. This helps in maintaining good accuracy of prediction, reducing inconsistencies, and enhancing the reliability of the model across all datasets. Here, the assimilation process identifies the relation between one feature and another, which helps recognize the exact patterns of incoming new patient features. Additionally, the introduced method uses a refined technique that lessens outliers during classification. As a bonus, the technique employs a robust education and instruction process that boosts recognition precision.

1.6 Research Contribution

The main research contribution is listed as follows.

1. The system's flexibility is maintained by utilizing data sampling techniques, namely Flock optimization in CNN, in the first layer to analyse the data association and accurately classify data characteristics.
2. Examine the underlying and most prominent characteristics using sampling and flock optimization methods. Efficiently extracting characteristics reduces the discrepancy between actual and forecasted glucose levels.
3. To maintain the system accuracy efficiency by using an Optimized Long Short-Term Memory network which improves the system training process and minimizes the time by updating the network parameters by correlated information from the previous stage. Hence the Data overfitting problem could be reduced. The optimization process minimizes the data assimilation and imbalance data issues successfully.

1.7 Thesis Organization

The thesis has been structured as follows:

Chapter one explains a detailed discussion of the problem statement, research contribution, scope of the study, and objective.

Chapter two discusses the basic discussion of diabetic disease, the type of disease, risk factors, diagnosis procedure, and treatment procedure is described. This chapter comprehensively reviews the literature study on diabetic disease is discussed. This chapter also covers the diabetic disease history and its impact on people's lives. This discusses the related works to investigate the earliest studies and shows the comprehensive research problem from the analysis. In addition, other comparative technologies are described to understand the efficiency of the existing diabetic prediction model.

Chapter three briefly discusses the research methodology, like Sampling Flock Optimization Assimilated Deep Learning Model.

Chapter four discusses the comparative study of the introduced diabetic prediction methods using the respective experimental analysis. This chapter also explains the data analysis and the proposed model's effectiveness.

Chapter five sums up the scope of the thesis and the Conclusions for future work in the research area.

1.8 Summary

Thus, the chapter discusses the introduction and basic foundation in the field of diabetic detection process. This begins by describing the characteristics and significance of diabetes illness, setting a backdrop for the ensuing research. The study attempt is motivated by the necessity of predicting diabetes outcomes. Subsequently, a thorough investigation of the background of diabetes diseases is carried out. This comprises a comprehensive inquiry that determines crucial elements, such as the classification of the problem, research goals, extent, and the expected impact of the study on the current corpus of knowledge. The chapter attempts to establish a comprehensive and precise framework for future research endeavours by exploring the beginnings of diabetes diseases. The formulation of research objectives is a crucial phase in this process. The objectives are carefully designed to target the recognized issues in the field of diabetes prediction. The primary objective is to improve the precision of diabetes prediction by employing a focused and methodical strategy based on the concerns outlined in the problem definition. A thesis outline has been designed in the concluding section of the chapter. The chapter attempts to enhance the overall coherence and efficacy of the research study by creating a well-structured framework for the thesis. This will ensure that each part of the study corresponds with the established objectives and makes a relevant contribution to the area of diabetes prediction. This chapter establishes the foundation for the whole research project by clarifying the issue environment, defining objectives, and outlining the path for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Diabetic (An et al., 2020) is the eighth leading dangerous and death-causing disease. In the United States, 37.3 million people suffer from diabetes. Diabetic disease is interconnected with other body parts. This causes severe complications (Sharma et al., 2022), such as stroke, heart disease, kidney failure, blindness, and lower limb amputations (Mandal et al., 2021). This also creates cancer, hearing loss, dementia, urinary incontinence, and erectile dysfunction. Diabetic disease occurs due to improper production or a high blood glucose or sugar level (Awuchi et al., 2022). According to the sugar level, diabetic disease has been categorized into Type 1, Type 2, Prediabetes, and Gestational diabetics (Awuchi et al., 2020). The World Health Organization (WHO) (Garg et al., 2020) report clearly states that 6% of adults have type 1 diabetes. In the U.S., 90 to 95% of adults have Type 2 diabetic disease, which gradually increases in children and adolescents. High-risk factors for type 2 diabetes increase the likelihood of developing prediabetes. Prediabetes infect around 96 million adults in the United States. Finally, gestational diabetic diseases have occurred in pregnant women. The diabetic infection rate rose from 108 million (1980) to 422 million (2014) (Hajam et al., 2023). The rate of increase in this illness is still greater in inadequate and moderate-income countries than in high-income regions.

According to the WHO report, from 2000 to 2019, the diabetic mortality rate increased to 3%. In 2019, 2 million deaths were recorded due to diabetic-related kidney disease. In 2014, 8.5% of people aged 18 and up had diabetes (Arokiasamy et al., 2021). About half of the 1.5 million deaths attributed to diabetes this year were in adults younger than 70. High blood sugar contributed to about 20% of all fatalities from heart disease (Wheeler et al., 2021) and was responsible for an additional 460 000 fatalities from kidney disease. When considering age, the death rate from diabetes rose by 3% between 2000 and 2019. The death toll from diabetes increased by 13% in poor and medium-income countries. By 2019, 22% fewer people aged 30–70 worldwide will perish from cancer, prolonged lower breathing disorders, cardiovascular disease, or diabetes than in 2000.

The WHO aims to support and initiate countermeasures to avoid and prevent diabetic disease and complications in undeveloped and developing countries. The WHO provides prevention guidelines to reduce the non-communicable disease called diabetes (Basu et al., 2020). In addition, WHO have certain standards, norms, and awareness programs to reduce the diabetic infection rate. The WHO created the awareness program on November 14 (Barber et al., 2022), and declared as the "World Diabetic Day." In addition, the "Global Report on Diabetes" is announced with recommendations, management procedures (Barreto et al., 2019), and prevention measures to reduce the mortality rate.

The World Health Organization (WHO) established the Global Diabetes Compact in April 2021 (Barke et al., 2022), a worldwide scheme to improve diabetes anticipation and care in developing and undeveloped countries. The compact covers entire

stakeholders to minimize the diabetic infection rate by improving the treatment quality and patient care. In May 2021, the WHO declared the Health Assembly to improve diabetic prevention and control recommendations. Increased availability of insulin, promotion of convergence, and harmonization of regulatory requirements for insulin and other diabetes-related therapies and health goods are all suggested. Five goals for diabetes care and treatment were approved by the World Health Assembly in 2019 (Poppin et al., 2019). Therefore, the World Health Organization (WHO) advises persons with diabetes to eat healthily, exercise regularly, keep a healthy weight, take the prescribed medications, and not smoke.

As explained in Chapter 1, diabetes is one of the most severe diseases for men and women (Iacopi et al., 2023). The improper insulin segregation reduces the body mechanism and causes the diabetic infection rate. Automatic diabetic detection techniques are being developed because of the significant risk factors connected with diabetes and its potential effects (Kakitapalli et al., 2021). Technology is used in these systems to identify those at risk for those newly diagnosed with diabetes. This technology allows for more targeted treatment and potentially reduces the cost of diabetic care. Data mining and machine learning methods are used by the detection system to increase the detection success rate. Furthermore, by utilizing data mining methods and machine learning methodologies in diabetes prediction, healthcare providers and researchers may leverage the power of data for improved accuracy and timeliness in identifying those at risk for developing diabetes. Improved patient outcomes and more efficient preventative measures are a direct result of the increased detection rate made possible by these technologies through the exposure of previously hidden patterns, the selection of pertinent features, the individualization of forecasts,

the adaptation to novel information, and the facilitation of real-time monitoring. The learning approaches analyse the input features and improve the overall recognition rate. The detection process uses various data that predict the changes in every data. However, the existing systems consume high computation complexity while analysing large volumes of data. High computational complexity and managing large datasets are essential for accurate diabetes prediction, yet these factors present substantial hurdles in real-time applications like ubiquitous health monitoring devices. Increased costs, longer computation times, scaling difficulties, inefficient resource use, and limitations on real-time applications are all examples of these problems. Existing systems necessitate powerful processing resources and specialized hardware, increasing costs and changing existing infrastructure. Delays in receiving results can also be caused by the intricacy of computations and analytical procedures, which is especially problematic in healthcare situations where rapid intervention is needed. The difficulties in the diabetic detection process are minimized with the help of optimized learning techniques (Saha et al., 2022), such as deep learning concepts and meta-heuristics optimization techniques. Hence, this chapter discusses introducing diabetic disease and its respective risk factors. In addition, various researchers' methodologies are described to understand the diabetic detection process.

2.2 Diabetic Disease

Diabetic disease is widespread due to blood sugar and high blood glucose. The human body receives energy from insulin, which is obtained from the glucose resource (Boucsein et al., 2021). The insulins are obtained from the pancreas, the main reason for the entire body's function. If the patients infected by diabetes-generated glucose

never reach the body cells, that will impact the body function. From the report of WHO, 30.2 million people were infected due to diabetic disease (Geraghty et al., 2019). As said, diabetes is the main reason for several diseases such as cardiovascular problems, being overweight, heart attack, and stroke. There are many subtypes (Khan et al., 2019) of diabetes, but the very common are Type 1 and Type 2 diabetes, both of which can develop during pregnancy.

Table 2.1: Characteristics of Various Types of Diabetes

Type of Diabetes	Description	Main Characteristics
Type 1	An autoimmune disease that targets the pancreas' insulin-making cells	Onset usually occurs in early childhood
Type 2	Inadequate insulin secretion and insulin resistance	Obesity is a common cause, and affects many adults.
Gestational	Insulin resistance-induced gestational hyperglycaemia	Usually occurs in the second or third trimester
Prediabetes	High blood sugar levels, or hyperglycaemia, are precursors to Type 2 diabetes.	Possibility of developing Type 2 diabetes
Monogenic	Caused by an isolated genetic change that decreases insulin production	Conditions like MODY and neonatal diabetes are extreme rarities.
Secondary	Stems from pre-existing disorders or medication use	Result of additional medical issues or medications
Drug-Induced	Caused by drugs that interfere with insulin's ability to work	Corticosteroids, antipsychotic drugs, etc.

Once the patient is affected by the diabetic disease, individuals experience the following symptoms (Wicaksana et al., 2020; Kumari et al., 2022; Yang et al., 2022) frequent thirst and hunger, frequent urination, improper weight loss and gain, irritability, fatigue, blurred vision, nausea, sluggish wounds healing and etc. A combination of genetic, environmental, and lifestyle factors influences diabetes. Individuals are more likely to develop diabetes if patient have these risk factors. These risk factors are essential for early detection, prevention, and efficient treatment. Key diabetic risk factors are obesity, family history, fat distribution, smoking, ethnicity and high cholesterol (Liccardo et al., 2019, Bansode and Prasad, 2022; Liang et al., 2023).

Long-term diabetics can develop the following complications (Ornoy et al., 2021; Sacchetta et al., 2022; Linn et al., 2023) from the condition. The above discussions show diabetic patients have severe symptoms, complications, and risk factors. Therefore, the patient requires an earlier detection system to predict the diabetic disease before reaches deadly complications. The clinicians suggest a self-monitoring process to identify the diabetic disease earlier. The following sections cover each of the processes that make up the monitoring process.

2.3 Diabetic Prediction models

2.3.1 Diabetic Self-Monitoring Process

The researchers and clinicians frequently reported that self-assessment (Du et al., 2020) is one of the important factors. Blood glucose levels must be checked during the assessment to control dinner planning, drug consumption, and physical movement. Self-Observing Blood Glucose (SMBG) (Pluta et al., 2021) is utilized in the assessment process to identify the blood glucose concentration. The test utilizes the test strip in which a small amount of blood is applied to predict the diabetic influence in the patient's body. During the analysis, the SMBG machine has certain conditions for effective results. The patient should ensure dry hands before handling the test strips and not use them more than once. During the testing, lapse data must be checked, and verify the machine should work properly. Once the patient is interested in using the machine or strip, the products must be stored in the cool territory, and the result should be discussed with the authorized doctors. After making the initial arrangement of the machine and strips, the patient must prick the skin, and a blood drip must be applied

to predict diabetes. The strips must be cleaned with warm water, and necessary actions must be taken. During the test, the patient must use the little finger, ring finger, and middle finger to drip the blood. After taking the test, the sharp items are disposed of carefully to minimize the risk factors. The self-assessment process initiated in the following situation is defined in Table 2.2.

Table 2.2: Frequency of Diabetic Disease Self-Assessment

Circumstance	Self-Assessment Recommendations
The patient who needs four or more insulin injections every day to control the respective condition	A minimum of four measurements of glucose in the blood per day (before meals, after meals (2 hours), risk of hypoglycaemia, and bedtime)
An individual who has been diagnosed with type 2 diabetes and is now being managed medically with once-daily insulin injections and oral hypoglycaemic agents	At least one glucose reading should be taken daily, ideally more than one (fasting, preceding meals, 2 hours during meals, or before night).
Diabetic taking insulin secretagogues for type 2 diabetes	At the first hint of low blood sugar (hypoglycaemia),
An individual with type 2 diabetes who is not at risk for hypoglycaemic or other complications due to the treatment regimen or lifestyle	In most cases, this is not essential, although certain exceptions exist.

According to the circumstance, self-assessment is performed, and the decision is taken to avoid the risk factor. The above analysis clearly shows that clinical systems require earlier detection procedures to improve the recognition rate. As a result, data mining and techniques for machine learning are used to develop an automated system for identifying diabetes. The detailed working process of the diabetic prediction model is described as follows.

2.3.2 Earlier Diabetic Prediction model

The self-assessment process is used to predict diabetic disease manually. However, most people fail to monitor the health at the correct time, creating complexities in

patient health. People with fewer symptoms are difficult to identify in the earlier stage, which maximizes the overall diabetic mortality rate. Based on data mining and machine learning methods, scientists have developed an automated system for predicting diabetes (Varma et al., 2019). The earlier detection system has several phases, which are designated in Figure 2.1.

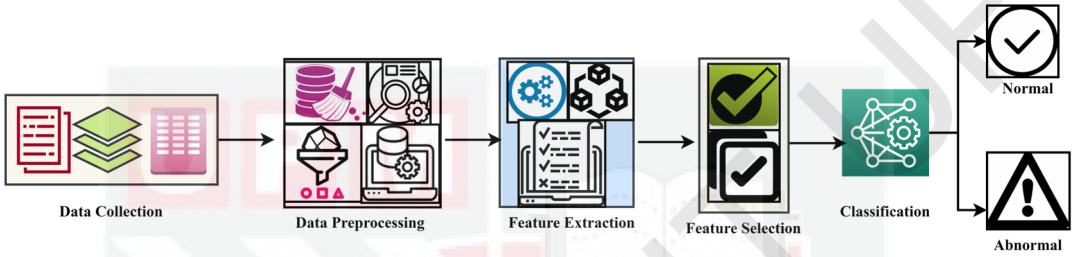


Figure 2.1: General Diabetic Prediction Process

Figure 2.1 illustrates the overall diabetic prediction process. Several processing phases make up the earlier detection system (David et al., 2020): acquisition of data, pre-processing, feature deduction, selection, and classification. Every step uses different processing methods, procedures, functions, and learning rules to improve diabetic detection efficiency.

Diabetic Data Collection

The evaluation of programs for the management and prevention of diabetes requires the collection of relevant data. Diabetes programs will collect data on diabetic patients and may employ various data collection tactics (Kamalraj et al., 2021). The types of data that will be obtained and the approach that will be used to collect those data will be determined by the specific assessment queries to be investigated and the assessment to be carried out. Most assessments use qualitative and quantitative data sources

(Aguinis et al., 2019). Complete solutions to evaluation issues may be found using a mixed-methods strategy.

The program development or implementation stage also influences the data collection tactics, which can vary. For instance, process reviews consider how the diabetes program is carried out. Data collection for process evaluations may focus on the following: participant engagement, staff and partner participation, program activity implementation, or the accessibility and usefulness of program resources. Information on program participation and execution and data on individual participants' perspectives and degrees of satisfaction can be gathered through process evaluations.

Different types of data are essential for conducting outcome evaluations, which analyse the results of diabetes treatment. The information collection for outcome evaluations may focus primarily on observing how the participants' skills, mind sets, and actions shift over time. Data of a quantitative (or numerical) nature and qualitative (or descriptive) nature are examples of information that could be acquired for outcome evaluations. Strategies for collecting quantitative data include introducing surveys or questionnaires that traditional channels like the telephone, postal service, or the World Wide Web can distribute. Quantitative data can also be gathered by mining already-existing resources like electronic health records and other clinical datasets. Qualitative data can be collected in various ways, including through conversations, focus groups, and open-ended questionnaires. Observational data, either quantitative or qualitative, is gathered using defined techniques to record actions, circumstances, and events.

Diabetic Data Pre-processing

The second phase of earlier detection is data pre-processing (Ali et al., 2020), which helps to remove the noise from the collected patient data. Data preparation is transforming raw data into a format that can be used and understood by a wide audience. Real-world data, often known as "raw data," often suffers from poor formatting, typos, and missing information because humans compiled the data. Pre-processing the data helps solve these problems and produce complete and effective datasets for data analysis. Everyone is aware that a database is a compilation of individual data points. Observations, data samples, occurrences, and recordings are different names for data points. Each sample's various qualities, often referred to as features or attributes, describe the samples. Developing models properly using these attributes is impossible without first pre-processing the data. During the data collection process, there is potential for great trouble. Therefore, data aggregation is performed for multiple data sources, which could result in a mismatch between data types like integer and float. Before processing the data, there must be four characteristics to improve the overall data analysis accuracy.

- The term "accuracy" refers to the state in which the knowledge possessed by a person is accurate. Errors, out-of-date information, and redundancies may impact the correctness of a dataset.
- Maintaining consistency requires ensuring that the data are free of any inconsistencies.
- Integrity: The dataset shouldn't have any fields that are missing data or that are incomplete, and shouldn't have any empty fields either. Due to this quality, data scientists can conduct precise analyses who have access to an all-encompassing picture of the data's circumstances.
- Validity is determined by the appropriateness of data samples in a dataset, including the type, adherence to a predetermined range, and correct format. Invalid datasets are challenging to organize and evaluate in a meaningful way.

- The data should be gathered when the event occurs to ensure maximum timeliness. Each dataset loses some precision and use as time passes since datasets are no longer accurately represents the current reality. As a result, one essential feature of high-quality data is the degree to which is both timely and pertinent.

Feature Extraction

The third step of this work is feature extraction (Maddumala et al. 2020), in which useful details are derived from the noise-removed data. The feature extraction states the practice of extracting valuable information from the information on diabetes obtained. The retrieved features are diabetic disease-linked patterns utilized to investigate the deviation of the typical diabetic structure. There are a few statistical derivations (Pooja Maule et al.,2015) of diabetic features that are listed as follows. There are a few characteristics that are present in diabetic extraction. These features are enumerated in Table 2.3.

Table 2.3: Feature List

Feature	Associated Formula
Entropy	$\sum_{i,j=0}^{n-1} -\ln(P_{ij})\, P_{ij}$
Correlation	$\sum_{i,j=0}^{n-1} P_{ij}\, \frac{(i-\mu)(j-\mu)}{\sigma^2}$
Energy	$\sum_{i,j=0}^{n-1} (P_{ij})^2$
Variance	$\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} (i-\mu)^2 . p(i,j)$
Mean	$\sum_{i=0}^{2(n-1)} i . p_{x+y}(i)$

According to the statistical characteristics discussed above, each patient has various pieces of information that are challenging to comprehend and take up more time and

complexity. Among the features that have been identified (Hina and Saadeh, 2022; Mengarelli et al., 2023) are:

- Entropy: The entropy in a data set indicates how chaotic or uncertain the data is. Entropy can be used to infer the degree of instability or unpredictability in diabetes-related characteristics, which may then reflect outliers or anomalies.
- Correlation: The correlation between the various factors in the diabetic data that examines the connections between variables can help explain the aetiology, progression, and management of diabetes.
- Energy: The energy of a piece of data is a metric that quantifies how prominent a given feature and that can be useful for drawing attention to key features of diabetes.
- Variance: The variance measures how far off individual data points are from one another. In diabetes, variance estimation can shed light on how much room there is for change in various observable characteristics.
- Mean: The mean, or average, of a set of numbers represents the most typical value within that set. The average values for diabetes-related characteristics can be gleaned by calculating the respective means.
- These features, taken as a whole, form the basis for feature extraction in diabetes studies. Researchers want to better understand the complex illness landscape by carefully examining and quantifying these characteristics using a variety of statistical derivations. In addition, the procedure of picking attributes to improve diabetes sickness identification could be enhanced.

Feature Selection

The collected dataset information is processed by a feature extraction method that derives the various features (Hegde et al., 2019). The dataset's many features introduce complexity and time in computing. The information prioritized from the retrieved features to enhance the data analysis procedure. Selecting the important feature, characteristics, or parameters is defined as feature selection. Reducing the number of input variables in a model can be accomplished through feature selection (Maniruzzaman et al., 2020), which involves selecting only the data relevant to the model and discarding irrelevant information. Due to which machine learning model's

features are automatically chosen based on solving the problem. This process is known as feature selection. To accomplish this, researchers will either include or omit significant features without making any changes. This helps cut down on the noise in our data and diminishes the size of the input data. The feature selection process is divided into two models such as supervised and unsupervised models. Selecting features that use the output label class is known as “supervised feature selection” in the framework of supervised models. This process uses the goal variables to find other ones that may help the model perform better. Unsupervised Models: Unsupervised feature selection is a method that does not require the outputs label class to select features. This method is used with unsupervised models. The feature selection categorization is illustrated in Figure 2.2.

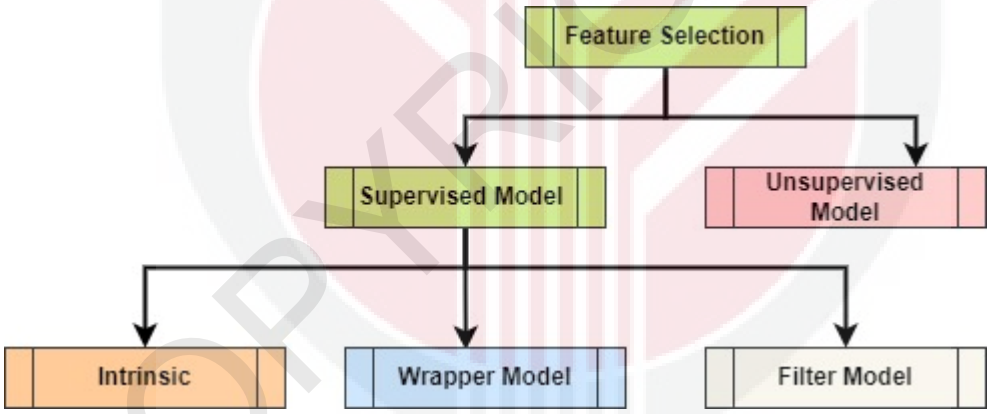


Figure 2.2: Taxonomy of Feature Selection

Figure 2.2 illustrates the categorization of the feature selection model. The supervised learning model is divided into three types such as intrinsic, wrapper model, and filter model (Brownlee et al., 2019; Homayouni et al., 2022).

Filter Model

The filter model selects the features according to the relationship and correlation. According to the feature correlation (positive or negative), the irrelevant features are dropped from the feature space. The filtering process related feature selection process is illustrated in Figure 2.3. In machine learning and data analysis, the supervised feature selection filter model cherry-picks important features from a dataset before feeding into a learning algorithm. This model ranks features based on criteria like correlation, statistical measurements, or knowledge gain without relying on the underlying learning method.

The term "filter" describes this strategy since this eliminates low-priority data before learning begins. Features like pre-processing, its sovereignty from the learning algorithm, feature ranking, selection criteria, and dimensionality reduction are fundamental to the filter model (Kumar et al., 2023). The dataset's properties and the study's objectives will determine the criteria utilized for selection. Features that show large variances across classes or highly correlate with the target variable could be given extra weight in classification tasks. Selecting crucial features from the dataset to reduce dimensionality is the primary goal of the filter model. This may result in shorter training periods, more easily interpretable models, and enhanced generalization to novel data.

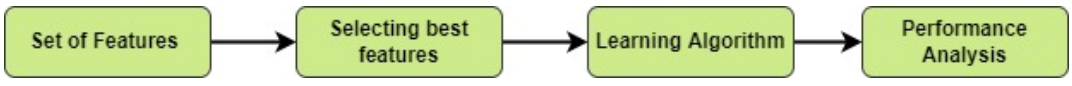


Figure 2.3: Filtering-based feature selection process

Wrapper Model

The wrapper strategy (Chong et al., 2021) is the subsequent model, and uses a training with subset breakdown of the extracted features. The training model output is utilized to create the sub-set features, and the subset process uses the greedy approach to evaluate the efficiency of the feature selection process. Wrapper models are supervised feature selection (Kanyongo and Ezugwu, 2023) that employ a dedicated learning algorithm to compare and contrast various feature subsets and identify the optimal combination for a given job. This model's distinguishing features include its iterative nature, algorithm-level evaluation, and performance measurements. The evaluation is typically performed using the same learning method as the target task, guaranteeing that the chosen feature subset is optimal for both. Feature subsets are evaluated based on performance criteria, including precision, recall, accuracy, and the F1-score, to determine suitability for a specific task.

Wrapper models get more computationally intensive when more feature subsets are employed for training and testing the learning algorithm. This difficulty level may prevent its use in high-level algorithms or with huge data sets. Overfitting is another concern, especially with smaller datasets. Recursive Feature Elimination (RFE) is a popular wrapper method that begins with all features before gradually eliminating the least important feature. Until either the target feature count is attained or the performance indicator stabilizes, the learning algorithm undergoes retraining after each feature is dropped.

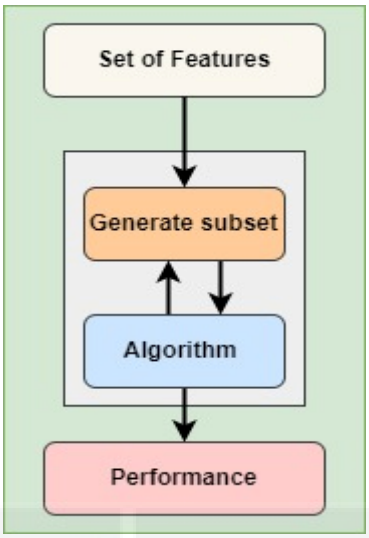


Figure 2.4: Wrapper Feature Selection Process

Intrinsic Model

The next model is the intrinsic approach that uses the wrapper and filter method (Mortazavi et al., 2021), which helps to generate the best feature set—the intrinsic-based feature selection process is illustrated in Figure 2.5. In machine learning and data analysis, the intrinsic approach to supervised feature selection is a technique that automatically discovers and picks important features based on the inherent value for a certain task (Cenitta et al., 2022). This strategy incorporates feature selection into the learning algorithm, prioritizing the selection of informative traits over the less useful counterparts to boost performance. Some of the most important features of the intrinsic model are its ability to adapt to different tasks, its seamless learning process, the prominence of changing features, increased efficiency, tamed complexity, and algorithm-level implementation. This attempts to determine which features significantly contribute to a given task's accuracy and generalization capacity by evaluating the value in light of the task's purpose, which may be classification, regression, or clustering. Intrinsic feature selection can boost the efficiency of machine learning algorithms like those that use decision trees or neural network topologies.

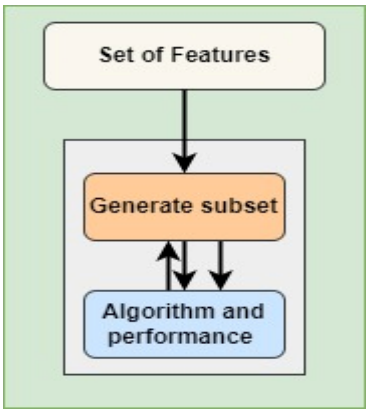


Figure 2.5: Intrinsic Feature Selection Model

Among the various feature selection approaches, the effective and optimized features are selected with the help of the optimization techniques. The best characteristics are chosen using several different optimization techniques, including those listed above, as well as Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Ant Bee Colony Optimization (ABC), and meta-heuristic approaches (Lalama et al., 2020). Methods for choosing useful features are used, which are then fed into categorization procedures. These methods use patterns acquired from labelled data to classify instances as normal or abnormal. Predictive models that are both accurate and meaningful can be achieved through efficient feature selection and suitable classification methods.

Classification

Classification (Shaban et al., 2020) is the final step in automatically recognizing diabetic disease. Recognizing the various feature-related categories is a necessary step in the process. According to different researchers, some examples of classification algorithms include the k-nearest neighbouring approach, decision trees, classification regression trees, support vector machines, and neural networks. Each algorithm

performs its tasks following the training notion in three ways: supervised, semi-supervised, and unsupervised. Gaining knowledge and experience through practical application yields positive outcomes for specific issues. Even though there are several different approaches to categorization, each algorithm has advantages and disadvantages. The algorithm has decided to categorize the features, taking into account the benefits successfully. Throughout each stage, researchers employ data mining and machine learning techniques to get better at spotting cases of diabetes. A few researchers' perspectives are presented here to contextualize the use of machine learning in diabetes diagnosis.

2.3.3 Literature Study of the various methods involved in the automatic detection system

This section discusses the various researcher's opinions, ideas, thoughts, and frameworks for creating an effective earlier diabetic detection system. Improved accuracy in diabetes prognosis and diagnosis was achieved with the help of a structure for machine learning (Olisah et al., 2022). Using Recursive Feature Elimination and Select Best, features are chosen using PIMA Indian and Medical City Hospital's laboratory datasets after missing value pre-processing. SVM, Random Forest, and Logistic Regression are only some of the classification algorithms used, along with cutting-edge missing value imputation methods. The F1-score, accuracy, precision, and recall improve the results. Accuracy on the PIMA Indian dataset is 99.01%; on the LMCH dataset, which the result is 99.57%, whereas the test prediction accuracy for Random Forest is 97.25%.

Accurate diabetes prediction was achieved by applying a fused machine-learning strategy (Ahmed et al., 2022). This method prioritized the analysis of symptoms and lifestyle characteristics in people with diabetes. The method includes a combination of SVM and ANN model training, fuzzy logic incorporation, and cloud storage for the resulting fused models. The results demonstrate superior performance than previously published approaches, with a high prediction accuracy of 94.87%. The fused machine learning model's real-time forecasts and excellent accuracy makes an intriguing option for early medical disease prediction.

The accuracy of diabetes prediction was increased using a labeled Bangladeshi diabetic dataset and a computerized classification pipeline (Dutta et al., 2022). The preprocessed ensemble model (DT + RF + XGB + LGB) achieved an accuracy of 0.735 and an AUC of 0.832. Early diabetes prediction was enhanced by using RF-based feature selection and statistical imputation. The research concluded that diabetes prediction might be improved using statistical restoration and RF-based feature selection methods.

Twelve persons with type 1 diabetes participated in a six-week longitudinal investigation that utilized continuous glucose measurement and an implantable sensor wristband to collect physiological data (Zhu et al., 2022). Hypo- and hyperglycemic episodes were significantly correlated with several physiological indicators in a mixed-effects logistic regression study. ARISES, a platform that uses a deep learning algorithm to predict glucose levels and hypo- and hyperglycemia, has shown promising results. The approach has a Matthews correlation value of 0.56 0.07 for identifying hypoglycemia and 0.70 0.05 for detecting hyperglycemia, with an RMSE of 35.28 5.77 mg/dl for an estimated horizon of 60 minutes.

Table 2.4: Summary of the Literature Review

Author & Year	Title	Method	Performance	Limitation	Datasets
Kezhi li et al. &2019	Convolutional Recurrent Neural Network for Glucose Prediction	Modified CRNN, LSTM	**Best result in the simulated dataset** than the real-time dataset	Performance in the prediction of hyperglycaemia degrades much faster than hyperglycaemia Hybrid model development can improve the performance	Real-time datasets of patients and simulated dataset
Kezhi li et al. &2020	GluNet: A Deep Learning Framework for Accurate Glucose Forecasting	GluNet: Personalized deep neural network with probabilistic distribution	The **RMSE of the Proposed model got good** results compared to SVR, LVX, NNPG, and ARX.	High-quality training data is required to retain prolonged accuracy results. The validation instances increase the time lag.	In-Silico, Clinical datasets
Martinsson et al & 2020	Blood glucose prediction with variance estimation using recurrent neural networks	Recurrent Neural Network Model	**Root-mean-squared error** (RMSE) for making a glucose prediction 30 and 60 minutes into the future, averaged across 100 randomly generated starting points for each patient and averaged for all patients, along with the corresponding means and standard deviations.	Scarce placement of wearable sensors data and ensures unsynchronized and dependent glucose values in the observation time. Therefore, the required data for processing and detecting glucose levels is limited.	Ohio T1DM
He et al. & 2020	Blood glucose concentration prediction based on kernel canonical correlation analysis with particle swarm optimization and error compensation	PSO-KCCA (Particle Swarm Optimization based Kernal Canonical Correlation Analysis)	The **RMSE values** for PSO-KCCA at PH = 5, 10, 15, 20, 25, and 30 minutes were 8.01, 11.98, 12.45, 13.23, 14.53, and 16.40 mg/dl, respectively. **R2 values** ranged from 0.95 to 0.98 to 0.97 to 0.98 to 0.97 on average.	The correlation between real-time and predicted glucose values shows up deviations. The deviations will be identified and used in training in different instances to reduce the errors in glucose level predictions.	Children's Diabetes Network Research Network (DirecNet) datasets
Wang et al. & 2020	Blood Glucose Prediction with VMD and LSTM Optimized by Improved Particle Swarm Optimization	(VMD-IPSO-LSTM) combining variational modal decomposition (VDM) and improved Particle swarm optimization optimizing Long short-term memory network (IPSO-LSTM)	The suggested VMD-IPSO-LSTM model generated **accurate** 30-minute, 45-minute, and 60-minute forecasts.	The execution complexity of the proposed method is high due to the assimilation of hard-level optimization and learning techniques. Multi-information consideration of the subject is not accounted for, which would improve the prediction's accuracy.	CGM data
H. S. Baruah et al., 2020	A Feature Selection Method using PSO-MI	**PSO-MI** (Particle Swarm Optimization with Mutual Information)	Threshold-independent values are utilized in this optimization technique to select the **relevant feature and remove irrelevant data**.	High dimensional dataset consumes more time and complexity	Spectrometer and NASA dataset
A. U. Rehman et al., 2020	Multi-Cluster Jumping Particle Swarm Optimization for Fast Convergence	**MCJ-PSO** (Multi-Cluster Jumping PSO)	MCJ-PSO **performs significantly** better than the existing systems while selecting the features from the dataset.	The research solutions are not guaranteed the exact solution, and the jumping behaviour cannot reach the exact solution. Accuracy is negligible for fast convergence.	CEC10 dataset

Table 2.4: Continued

A. Karale, et al., 2020	A Hybrid PSO-MiLOF Approach for Outlier Detection in Streaming Data	**Hybrid PSO-MiLOF Approach** (Hybrid Particle Swarm Optimized –Memory Efficient Local Outlier Factor)	Effectively **detects the outliers from stream data by resolving the problem** of finding the local outliers in stream data.	Reliability and anomaly detection with maximum accuracy is still one of the issues.	Real-time dataset
Zak M., Krzyżak A. (2020)	Classification of Lung Diseases Using Deep Learning Models.	**DL** (Deep Learning Model)	Predicting lung cancer disease by medical **data scarcity issues**.	This requires optimized techniques to improve the overall disease detection system.	ImageNet dataset
N. Fatima, et al., 2020	Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis	Data mining and Machine Learning Techniques	Comparative analysis of various machine learning techniques to predict the breast cancer **maximum accuracy**	An imbalanced number of breast cancer images examined the affected region and limited datasets.	Several real-time datasets
X. Zhang et al. 2020	Deep Learning-Based Analysis of Breast Cancer Using Advanced Ensemble Classifier and Linear Discriminant Analysis	**AEC-LDA** (Advanced Ensemble Classifier and Linear discriminate Analysis)	Predicting breast cancer by analysing medical gene expression data with **maximum accuracy**	This is generalized for particular data and should be applied to the publicly available dataset.	GEO dataset
M. Byra et al., 2020	Early Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer Sonography Using Siamese Convolutional Neural Networks	**CNN** (Convolution Neural Networks)	Applying Siamese convolution network to predict the Neoadjuvant Chemotherapy in Breast Cancer Sonography w**ith minimum error rate.**	Difficult to apply to large patient groups.	ImageNet and Breast Mass dataset
D. Karimi, et al., 2020	Deep Learning-Based Gleason Grading of Prostate Cancer from Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation	**Hybrid method** (Convolution network and image augmentation with feature space model)	Improving prostate cancer recognition **accuracy** by using a hybrid deep learning model	Data augmentation in high dimensional space easily fails.	Histopathology Images
S. Park, et al., 2016	Intra- and Inter-Fractional Variation Prediction of Lung Tumours Using Fuzzy Deep Learning	**IIFDL** (intra- and inter-fraction fuzzy deep learning)	Predicting intra and interfraction variations while analysing lung tumour with minimum **error**	The need to improve prediction accuracy and correlation between the tumour location in the lung is failure to detect.	Georgetown University medical centre using the CyberKnife Synchrony (Accuracy Inc. Sunnyvale, CA) treatment facility

Table 2.4: Continued

Q. Liang et al., 2019	Weakly Supervised Biomedical Image Segmentation by Reiterative Learning	**RL** (Reiterative Learning)	Segmenting gastric cancer region using reiterative learning process. This aims to **maximize the converging speed and minimize the loss function.**	Fail to recognize small cancer tissues; overfitting may occur during the reiterative learning process.	e 2017 China Big Data and Artificial Intelligence Innovation and Entrepreneurship Competition.
C. Du et al., 2019	Reconstructing Perceived Images from Human Brain Activities with Bayesian Deep Multiview Learning	**BDML** (Bayesian Deep Multiview Learning)	Analysing brain activity images for reconstructing the perceived images with **maximum accuracy.**	Human visual experiences are dynamic, which is difficult to predict	Binary Contrast Patterns, Handwritten Digits, and Handwritten Characters
V. Kumar et al., 2021	Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes,"	**Hybrid** (ensembling and deep learning machine learning technique)	Resolving **co-morbidity recognition** from patient's clinical records by using the hybrid machine-learning technique	Multi-label class classification problems should address	n2c2 natural language processing research dataset
M. R. Ahmed et al., 2019	Neuroimaging and Machine Learning for Dementia Diagnosis: Recent Advancements and Future Prospects,	**Machine Learning techniques** (Multimodal imaging analysis deep learning approaches)	Improving dementia prediction **accuracy** by using various machine-learning techniques	Need to improve the classification accuracy further.	MovieLens 10M and MovieLens 20M
D. Ahmedt-Aristizabal et al., 2021	Identification of Children at Risk of Schizophrenia via Deep Learning and EEG Responses	**DL** (Deep learning)	Analysing EEG Signals to predict the children's schizophrenia **risk rate**	external memory modules capable of mapping relationships across participants	Experimentally recorded
H. Ali, et al., 2020	A Deep Learning Pipeline for Identification of Motor Units in Musculoskeletal Ultrasound	**CNN** (Convolution Neural Network)	This has significant **performance improvement** while predicting motor units in the ultrasound images	learning the data is a more generalized approach	TIMIT dataset
Y. Chen, et al., 2020	Vertebrae Identification and Localization Utilizing Fully Convolutional Networks and a Hidden Markov Model	**Hybrid** (fully convolution and Hidden Markov Model)	**Minimize the complexity** while identifying the vertebrae in images	Lack of flexibility	MICCAI 2014 Computational Challenge on Vertebrae Localization and Identification

Table 2.4: Continued

T. D. Pham et al., 2010	Classification of short time series in early Parkinson' s disease with deep learning of fuzzy recurrence plots	**LSTM** (Long Short Term Memory Network)	Analysing short time series Parkinson's disease dataset and improving **recognition accuracy**	The reliability and efficiency of the system should be improved	neuroQWERTY MIT-CSXPD database
H. Nahas, et al., 2020	A Deep Learning Approach to Resolve Aliasing Artifacts in Ultrasound Colour Flow Imaging (CFI)	**CNN** (Convolution Neural Networks)	resolve **CFI aliasing artifacts** that arise from single- and double-aliasing scenarios using deep learning networks	Potentially improve the CFI in vascular diagnostics	Data collected from the scanner
T. Tuncer, et al., 2020	Novel Multi-Centre and Threshold Ternary Pattern-Based Method for Disease Detection Method Using Voice	**MCMTTP** (multi-centre and multi-threshold based ternary pattern)	Compared to existing techniques, fused features are **more useful** for detecting the disease	Accuracy should be improved	Saarbruecken Voice Database
S. Kaur et al,, 2020	Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives	**AI** (Artificial Intelligence techniques)	Predicting various clinical cancer to **improve the overall recognition accuracy**	Deviation should be reduced	Publicly available datasets
Ezzat et al., 2020	An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization	**Hybrid** (Convolution network with a gravitational search algorithm	Able to classify the covid 19 disease with **up to95% accuracy**	The difference between normal and covid 19 should be addressed.	Binary COVID-19 dataset.
Kaur et al., 2020	Hyper-parameter optimization of a deep learning model for the prediction of Parkinson's disease	**Hybrid (**deep learning and grid search optimization)	This shows **better performance** than state-of-the-art fine-tuning deep learning model	corresponding variant of grid search optimization on GPU should be considered	Wisconsin breast cancer dataset
Ramesh et al., 2020	Recognition and classification of paddy leaf diseases using Optimized Deep Neural network with Jaya algorithm	**Hybrid** (deep network and Jaya optimization algorithm)	This model proposed a new automatic hybridized model to predict the disease from paddy with **high accuracy.**	Need to minimize the false positive rate	Rice plant images
Cherian, et al., 2020	Weight-optimized neural network for heart disease prediction using hybrid lion plus particle swarm algorithm	**Hybrid** (hybrid lion plus particle swarm algorithm)	This shows **better performance** than other methods; This also shares the weight and optimizes the network weight while classifying heart disease	Completion time should be minimum	UCI repository heart datasets

Table 2.4: Continued

de La Torre et al.,2020	A deep learning interpretable classifier for diabetic retinopathy disease grading	**Pixel-wise score propagation model**	Examining the pixel contribution towards the diabetic retinopathy **prediction accuracy**	Lack of RD images, disease region prediction should be improved	Messidor-2 Dataset.
Fouad et al., 2020	Analysing patient health information based on IoT sensors with AI for improving patient assistance in the future direction	**IGDBN** (iterative golden section optimized deep belief neural network)	Improving the patient details **access accuracy** and health disease examination process	Accuracy needs to be improved	IoT sensor devices are utilized to collect the data
AlZubi, et al, 2020	Deep brain stimulation wearable IoT sensor device-based Parkinson brain disorder detection using heuristic tubu optimized sequence modular neural network	**Hybrid** (deep brain stimulation and heuristic tubu optimized sequence modular neural network)	recognize the changes in brain function with high speed that reduces the **prediction delay**, which helps to improve the Parkinson's disease seriousness	Flexibility and accuracy should consider	Florida university collects
Agrawal S. et al., 2020	Genetically Optimized Deep Neural Learning for Breast Cancer Prediction	**GODNL** (Genetically Optimized Deep Neural Learning)	Analyzing datasets and predicting breast cancer with **minimum time**	Accuracy should be improved	Wisconsin Breast Cancer Dataset
Olisah et al., 2022	Diabetes mellitus prediction and diagnosis from a data pre-processing and machine learning perspective.	**Random Forest, SVM, and 2GDNN** (Twice-Growth Deep Neural Network)	The proposed 2GDNN model achieves **97.34% precision, 97.24% sensitivity, F1-score, 99.01% train accuracy, and 97.25% test accuracy** in evaluations on the PIMA Indian and LMCH diabetes datasets.	The accuracy and reliability of predictions should be enhanced.	PIMA Indian and LMCH diabetes datasets.
Ahmed et al., 2022	Prediction of Diabetes Empowered with Fused Machine Learning	**Fused Model** (SVM and ANN)	The results show a **high prediction accuracy of 94.87%.**	The accuracy rate should be increased.	UCI repository diabetes dataset.
Dutta et al., 2022	Early Prediction of Diabetes Using an Ensemble of Machine Learning Models	**Ensemble model** (DT + RF + XGB + LGB) with RF-based Feature Selection	Achieved the **highest accuracy of 0.735 and an AUC of 0.832.**	The rate of accuracy and AUC to be enhanced	Bangladeshi diabetes dataset
Zhu et al., 2022	Enhancing self-management in type 1 diabetes with wearables and deep learning	**ARISES** (Adaptive, Real-time, and Intelligent System to Enhance Self-care) with Deep Learning	The method has a root-mean-**squared error (RMSE)** of 35.28 5.77 mg/dl for a prediction horizon of 60 minutes and a Matthews correlation coefficient of 0.56 0.07 for detecting hypoglycaemia and 0.70 0.05 for detecting hyperglycaemia.	The six-week study duration might not capture longer-term variations in glucose dynamics.	Real-time data from continuous glucose monitoring (CGM)

According to the above researcher's opinion, the mind map is created to arrange the working process of the diabetic detection system, which is shown in Figure 2.6.

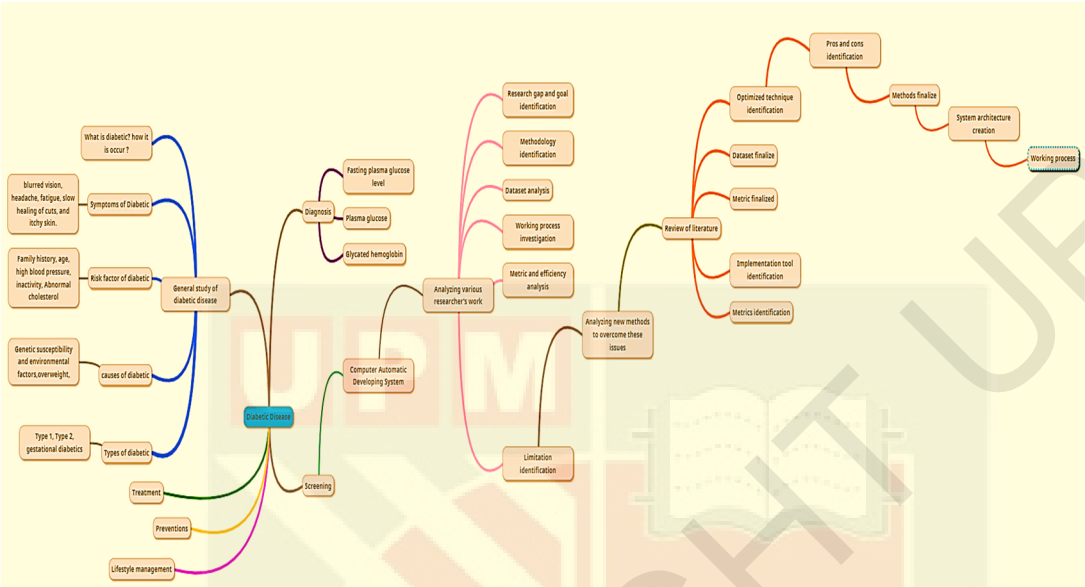


Figure 2.6: Mind Map Analysis of Diabetic Detection Process

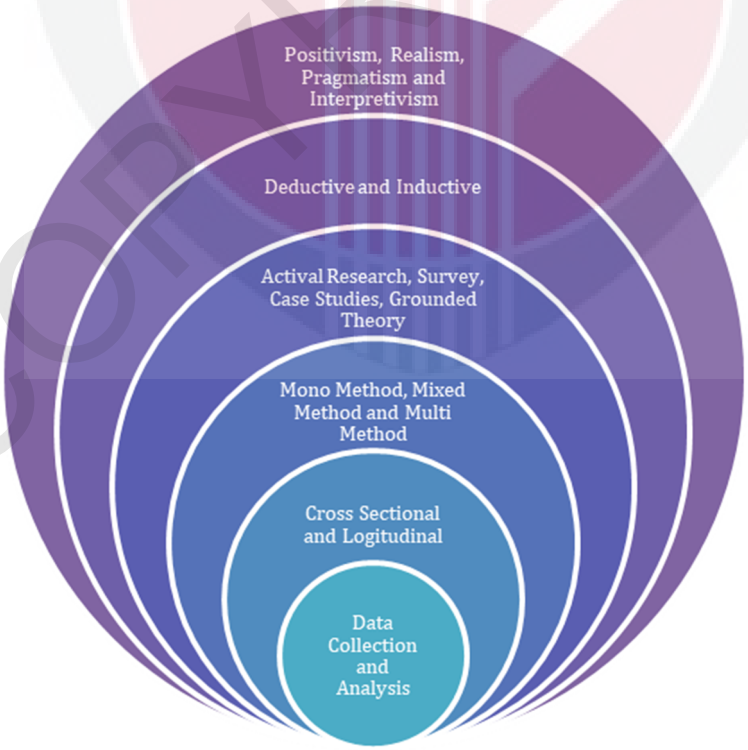


Figure 2.7: Onion Diagram of Diabetic Detection Process

2.3.4 Literature Review of Automatic Diabetic Detection Process

This article provides an overview of the various methods used to foresee the onset of diabetes. A powerful model for detecting diabetes was created by Ahmed Hamza Osman et al. (2022) using a hybrid of k-means clustering and support vector machines (SVM). Diabetic data is obtained from the UCI Pima Indians dataset, and the noise removal approach is used to eliminate any anomalies or missing information. The SVM is used to conduct the t-test after the noise-free data has been processed using the k-means method, which identifies the best features. The effectiveness of the system is assessed based on experimental results.

Kemal Polat et al.(2007) introduced an intelligent strategy for enhancing diabetes diagnosis accuracy using the Adaptive Neuro-Fuzzy Inference System (ANFIS) and Principal Component Analysis (PCA). PCA is specifically used to diminish the number of attributes in the diabetes dataset. The critical task of ANFIS is to create a classification model based on reduced attributes for early diabetes detection.

Mohammad Reza Daliri et al. (2012) employed the binary particle swarm optimization technique to investigate several medical diseases utilizing various data sets, such as the Pima Indian diabetes dataset, the heart database, the dermatology dataset, and the Wisconsin breast cancer dataset. The enhanced method evaluates the usefulness of four datasets and picks useful characteristics based on its position and velocity values. The chosen features aid in disease prediction without adding unnecessary complexity. The system's performance is then measured against the standard genetic feature selection method, with the results showing improved accuracy, information gain, and F-score value.

Naveena, S., & Bharathi, A. (2022) introduced moth flame related crow searching deep learning approach for designing diabetic detection and glucose level prediction model. The main intention of this work is to improve the diabetic prediction accuracy. During the analysis, UCI and PIMA dataset for exploring the diabetic details. The gathered information is processed with the help of Convolution Neural Networks (CNN) that predict the diabetic patients by addressing the multi-objective optimization problem. The network performance is fine-tuned with the help of the moth flame-related crow searching approach that predict the optimized parameters and reduces the redundant information.

Diabetes data from Urumqi, Xinjiang, China, was analyzed by Yuki Li et al. (2018) using data mining techniques such as SVM, integrated learning models, and decision trees. SMOTE, one of the effective feature selection algorithms, was first applied to data collected from Chinese individuals, together with an unbalanced technique that selects features such as body mass index (BMI), glucose level (glucose), and age (age is an important variable). The data miner above classifies the examined traits, reliably discriminating between healthy and diseased ones. Compared to traditional support vector machine classifiers, the SMOTE-based diabetes recognition system performs better. Knowledge and thoughts regarding the prognosis of diabetic illness appear to have been accumulated from the perspectives of various authors through various processing activities, as suggested in the debate.

The algorithm had Matthews correlation values of 0.56 0.07 and 0.70 0.05 for diagnosing hypoglycemia and hyperglycemia, respectively, and an RMSE of 35.28 5.77 mg/dl for a 60-minute prognostic horizon. The method presented by Bilal et al.

2022) uses preprocessing techniques and data augmentation to improve both image quality and quantity. The optic disc and retinal blood vessels are split using two U-Net models. After the data has been preprocessed, an identical hybrid CNN-SVD model is constructed, and Inception-V3 is used to extract discriminant features. Diabetic retinopathy (DR) can be diagnosed with this approach because this can spot micro aneurysms, hemorrhages, and exudates in the retina. The approach is evaluated on the state-of-the-art EyePACS-1, Messidor-2, and DIARETDB0 datasets. The results demonstrate that deep neural networks (DNN), particularly U-Net and CNN-SVD, greatly enhance the automated DR process of classification over conventional approaches. In addition to providing vital data for further research, the approach accurately segments the optical disc and blood vessels.

(Özbay, 2023) suggested a technique for identifying retinal lesions in retinal fundus pictures by employing the artificial bee colony (ABC) method for image segmentation. Threshold values are calculated from the image's histogram to enhance segmentation precision. To detect the different stages of diabetic retinopathy, a multi-layer active deep learning (ADL) system is created. The ADL-CNN architecture is developed. Key lesson features are identified, and regions of interest are segmented using a two-stage procedure involving training on pictures with basic accuracy labels and generating masks using the extracted retinal features. Classification accuracy, sensitivity, specificity, and the F-measure are the only statistical metrics used to evaluate performance. A classification precision of 99.66%, sensitivity of 93.76%, specificity of 96.71%, and F-measure of 94.58% are all impressive results for the ADL-CNN model on the EyePacs dataset.

A method was developed to improve the quality, eliminate noise, and standardize the format of retinal images from different sources (Menaouer et al., 2022). A deep convolutional neural network (CNN) is employed alongside two versions of the VGG network (VGG16 and VGG19) to improve feature extraction. Next, the model is built to identify and categorize diabetic retinopathy according to the degree of visual risk and retinal ischemia. The model is trained using the preprocessed dataset to understand patterns and features characteristic of various DR stages. These results demonstrate that the hybrid deep learning strategy outperforms prior methods for identifying and classifying diabetic retinopathy. Diabetic retinopathy may be detected and classified automatically according to visual risk using the proposed hybrid approach. The outcomes demonstrate a 90.60% precision, 95% recall, and 94% F1 score.

An approach was provided by (Mujeeb Rahman et al., 2022) to gather and preprocess digital fundus photos to make them appropriate for analysis. The SVM and a Deep Neural Network (DNN) are trained on the dataset. The next step is to undertake feature extraction and training to identify anomalies and trends that point to diabetic retinopathy. Pre-recorded digital fundus images can be used by SVM and DNN models to accurately predict the existence of diabetic retinopathy, according to the study. When applied to test data, AUC was 97.11% for the SVM model and 99.15% for the DNN model.

Medical imaging and records from patients with diabetes are collected, preprocessed, and fed into a Deep Belief Network (DBN) designed to predict and classify diabetes (Shahin et al., 2023). When training the DBN, need to consider features and patterns associated with diabetes progression and use the preprocessed diabetic dataset.

Accuracy and other metrics are used to judge how well the trained DBN model performs. This research proves that deep learning methods, particularly the Deep Belief Network (DBN), may be used to predict and classify the onset of Type 2 diabetes effectively. The accuracy of the suggested DBN method on the diabetic dataset was 81.25%.

Data collection, preprocessing, correcting data imbalances, and machine learning model selection were the main focuses of Uddin et al. (2023). The research proposed a collection of algorithms as follows: linear regression, logistic regression, k-nearest neighbor, naive bayes, random forest, support vector machine, and decision tree. The 2019 diabetes dataset and the Pima Indian dataset demonstrate the potential for the Random Forest algorithm to predict diabetes accurately. Using SMOTE to correct imbalanced datasets boosts the model's accuracy and decreases false-negative detections by a large margin. On the 2019 diabetes dataset, the Random Forest algorithm obtained 97% accuracy, while on the Pima Indian dataset, which reached 80% accuracy.

Exhaled breath samples are collected, analyzed for acetone concentration, and then used to create a sensing component with sensors and implement a deep learning algorithm (Bhaskar et al., 2023). Medical data analysis and prediction are two Convolutional Neural Network (CNN) applications. To solve this problem, need to create a unique deep hybrid correlational neural network (CORNN). The model is trained and tested using information on the diabetes status of the subjects in the breath samples. Precision, dependability, recall, and the F1 score are just criteria for evaluating the model's performance. The findings of this study demonstrate that the

concentration of acetone in exhaled breath is a more accurate biomarker for the diagnosis of type 2 diabetes than current approaches. Developing a novel sensor module and deep-learning neural network model has allowed for the non-invasive and highly accurate diagnosis of diabetes. The evaluation shows a predictive accuracy for diabetes of 98.02%.

This section discusses several methods for identifying and classifying diabetes and presents novel tactics. Overall, these studies add to the growing body of evidence for identifying, categorizing, and predicting diabetes, which could eventually lead to a more accurate early diagnosis and better management of the condition.

2.3.5 Review of Preprocessing Phase

This part aims to provide a comprehensive understanding of the method for eliminating noise by considering the thoughts, viewpoints, and functions of various authors regarding the data preparation strategy. (2015) To quote: (Seyed Hossein Rasta et al.) A report was written to see how changes in contrast and brightness during processing affected the quality of images seen by the retina. The examined and assessed several light adjustments and differentiation upgrade procedures for shading retinal images to identify the best approach for optimal picture enhancement. To evaluate and pick the best illumination rectification method, empirically and externally dissected the corrected red and green segments of shading retinal images. The influence and transparency of the two complexity improvement algorithms were determined using a vessel division computation. Coefficients of variation were used for the quantitative assessment of light therapy devices. The lowest coefficients of variation were obtained

in the red spectrum when a solitary approach was utilized to measure base brightness through the central channel. The residual and homomorphic shifting strategies produced excellent results using the partitioning strategy because of the low variation coefficients. The distinction-restricted flexible histogram levelling increased the affectability of the vessel division computation by up to 5% with a comparable measure of precision.

Many methods for identifying micro aneurysms, haemorrhage, and exudates are being explored by Javeria Amin and colleagues (2016) to diagnose non-proliferative diabetic retinopathy. The use of vein identification devices to diagnose proliferative diabetic retinopathy is also being considered. The paper also reviews the authors' previous work on detecting diabetic retinopathy. Experts and specialists who operate in this field and need to do ongoing studies will benefit from this work.

The current frameworks for screening for diabetic retinopathy are discussed (Rahim S.S. et al., 2016), with an emphasis on maculopathy detection methods. The proposed medicinal choice emotionally supportive network is divided into four sections: image securing, image pre-processing (including the localization of four retinal components), highlight extraction, and the order of diabetic retinopathy and maculopathy. The proposed structure uses several methods for extracting elements from images, including the Circular Hough Transform and various "fluffy" picture processing algorithms. In addition, a new method of pinpointing the macula region to identify maculopathy is proposed. The study introduces a fresh online dataset and provides details on the data's accumulation, master determination technique, and advantages of other open-eye fundus image databases used to study diabetic retinopathy.

Subbuthai et al. (2015) An original formula eliminates the annoying pixel in the retina image. Since green-channel retinal vessels are more visually appealing, the suggested method, Extended Median Filter, is related to the green channel of the retina. The proposed extended middle channel is compared to the present standard middle channel using execution metrics, including peak signal-to-noise ratio, mean square error, and RMSE. Testing findings show that the suggested Extended Median Filter computation outperforms the conventional middle channel regarding noise concealment and detail protection.

The purpose of the study by Febrian et al. (2023), which compared the k-Nearest Neighbour (KNN) and Naive Bayes (NB) machine learning algorithms, was to predict diabetes better. Data pre-treatment was essential to ensuring accuracy and reliability during analysis. Preprocessing steps include cleaning, integration, reducing, and discretizing to get data suitable for analysis. Incomplete, wrong, inaccurate, and missing data are only some problems that can be fixed throughout the data-cleaning process. Integrating data involves converting the original data's measuring scale into a format that can be read by the analytic tool. By reducing the volume of data, this may save money, time, and storage space. Data transformation is all about transforming data into a comprehensible format using an analysis tool. The study ensured the dataset was ready for direct analysis without subsequent transformation, reduction, or integration by employing these data pre-treatment techniques, which included verifying the absence of missing values and addressing potentially incorrect zero values.

The study's intention by Howlader et al. (2022) was to classify patients with type 2 diabetes (T2D) and discover variables that separate distinct sub-types for prognosis and treatment using machine learning techniques. The data set included Pima Indian women over 21 living in or around Phoenix, Arizona, USA. Only 268 of the 768 patients had type 2 diabetes, whereas the other 500 did not. The dataset had no blanks but was cleaned up, and some outliers were removed. Many feature selection strategies were used to extract informative feature subsets to prepare the dataset for analysis. Following data preparation and feature selection, compared machine learning models by looking at how well each dataset performed. Accuracy, kappa statistics, the area under the receiver operating characteristic curve, specificity, sensitivity, and log loss were some of the evaluation measures used to compare the results.

2.3.6 Review of Feature Extraction Phase

The following phase is called feature derivation, and this involves several authors analysing perspectives to extract a variety of features from the information collected on patients. This section aims to familiarize the feature extraction process by discussing various approaches and points of view. (Ioannis Kavakiotis et al., 2017). The current research aims to conduct a systematic survey of how artificial intelligence (AI), information mining systems (IMSs), and instruments are being used to advance the researchers understanding of diabetes in four key areas: a) prediction and diagnosis; b) diabetic complications; c) genetic background and environment; and e) health care and management. The AI's calculations extended to many contexts. Overall, 85% were described using regulated learning methods, while 15% were described most precisely using unsupervised learning methods. Support vector machines (SVMs) are now the

most efficient and popular computation method. Most of the time, clinical datasets were mined for information. The title applications chosen for these papers speculate on the usefulness of eliminating profitable learning, which prompts new hypotheses that concentrate on a more profound grasp of DM and additional research.

(Saiprasad Ravishankar et al., 2009) present another limitation for the identification of the optic plate. In this method, first identify the important veins and then use the place where intersects to locate the approximate size of the optic circle. This method has its limitations. In addition to this, the use of shading attributes places further restrictions.

Additionally, that various morphological jobs can reliably distinguish several highlights, including veins, exudates, micro aneurysms, and haemorrhages. The calculation was run on a database of 516 images that varied in complexity, brightness, and disease stage; the results showed a 97.1% success rate for optic plate limitation, an affectability of 95.7% and a particularity of 94.2% for exudate recognition, and a location accuracy of 95.1% and a particularity of 90.5% for micro aneurysms and discharges, respectively. This contrasts favourably in every way with the currently available frameworks and guarantee the authentic transmission of these frameworks.

(M. Akshatha Rao et al., 2014) present a robotized analysis of DR that uses another method known as the Hurst Exponent to determine the Fractal Dimension (FD). Contrast, correlation, energy, homogeneity, and entropy highlight the picture's dark dimension co-event lattice. Below is the measured examination of DR and healthy retinopathy for several highlights that have been separated from one another.

Ophthalmologists can quickly examine DR based on visual principles once obtained the Power Spectrum, which is used to obtain information about the retina.

Swapna G et al. (2018) detail a strategy for ranking HRV signals from people with diabetes and those without using deep learning models. The complex, fleeting, and powerful highlights of the HRV data are removed by the Long Short-Term Memory (LSTM), convolutional neural system (CNN), and its mixes. These highlights are then sent to a support vector machine (SVM) to be ranked, and without applying SVM in researchers past work, got an outstanding improvement of 0.03% and 0.06% in CNN and CNN-LSTM design, respectively. This improvement was compared to previous work. The suggested grouping structure has the potential to assist doctors in the diagnosis of diabetes by utilizing ECG data with a very high degree of accuracy, which is 95.7%.

(Mishra, Sushruta, et al., 2016) Identifying infectious diseases is one of the application domains where AI instruments are achieving success. Diabetes is a disease that affects people worldwide and is one of the key factors that might lead to mortality. Due to the availability of vast amounts of medical information, there is a growing demand for exceptional learning tools that can assist medical professionals in correctly diagnosing diabetes. Using AI methodologies can help determine whether or not someone has diabetes, which appears to be a sensible feature of proficiency. Regardless, the material presented is repeated and noisy in its overall tone, which works against observing learning and serves as a good example. AI processes have attracted a significant amount of attention from the professionals who are responsible for converting such knowledge into useful learning. Using techniques that are channel based, this is

possible to eliminate more important data from massive amounts of recordings. In the context of diabetes infection, research analysis compares four separate channel-based component identification procedures (the Chi-square technique, the information gain approach, the cluster variation strategy, and the correlation technique). The presentation of the calculations was evaluated using RBF, IBK, and JRip, which are three different classifiers. According to the study, channel-based element choice strategies are superior to others in demonstrating learning calculations for accurate forecasting and diagnosing diabetic illness.

The prevalence of diabetes mellitus in Ethiopia was higher than expected by the International Diabetes Federation (IDFA), according to the work of Shiferaw Birhanu Aynalem et al. (2018). Related risk variables that may be altered were also identified in this study. This study is important to remember that screening for diabetes mellitus is especially important for people who have a high waist circumference, a history of smoking propensity, or hypertension. In this way, focusing on the avoidance technique of such modifiable hazard components may decrease the prevalence of diabetes mellitus.

(Okeke Stephen et al., 2019) present a convolutional neural system model developed from scratch without any external help for identifying pneumonia in a database of chest X-ray images. The convolutional neural system model is built without any preparation to remove highlights from a given chest X-ray image and arranged to decide whether an individual is infected with pneumonia. This contrasts with other strategies that depend entirely on exchange learning approaches or conventional high-quality procedures to achieve surprising order execution. These strategies are utilized to

accomplish surprising order execution. In contrast to these strategies, built a convolutional neural system model. The issues of reliability and interpretability frequently encountered when handling restorative symbolism may be made easier to manage with the help of this paradigm. The process is hard to acquire a lot of pneumonia datasets for this grouping task; consequently, few information increase calculations to increase the approval and arrangement precision of the CNN model, and accomplished surprising approval exactness. Unlike other deep learning characterization assignments with adequate picture vaults, acquiring a lot of pneumonia datasets for this grouping task is difficult.

The Tervaert neurotic characterization divides DN into four configurations for glomerular sores, and a distinct scoring system for cylindrical, interstitial, and vascular sores is presented in Chenyang Qi et al. (2017). The results were reported in nephrology. Essentiality and conflicts in conclusion and visualization exist in Tervaert's arrangement. This is because of the variety of diabetic nephropathy wounds and the sophistication of diagnostic tools. This review compiles the various implementations and assessments of Tervaert characterization and signals for renal biopsy because of late examinations suggested that this was necessary. Meanwhile, the audit discusses and deduces the situation where a regular DN is layered on top of a nondiabetic renal ailment (NDRD) and the differential conclusion with another nodular glomerulopathy. These topics pertain to a regular DN being superimposed with an NDRD.

A novel computation using a deep convolutional neural network (DCNN) is presented by Yun-Hui Li et al. (2019). In a major departure from standard DCNN practice, swap

out traditional max-pooling layers for partial max-pooling. To infer increasingly discriminative highlights for grouping, two of these DCNNs have been developed, each with a different number of layers. By familiarizing with the class's hidden dispersion limit, combine the DCNNs' output with the highlights from the image's metadata and train an SVM classifier. This study took advantage of Kaggle's publicly available DR discovery database. The constructed the model with 34,124 preparatory photographs, 1,000 approval pictures, and 53,572 testing pictures. The phases of DR are separated into five distinct classifications according to the DR classifier that has been suggested. Each classification is denoted by a whole number from zero to four. The results of the exploratory research indicate that the proposed technique has the potential to achieve an acknowledgment rate of up to 86.17%, which is higher than what has been lately reported in writing. In addition to developing an AI calculation, an application called "Profound Retina" is developed. When a normal person is provided with a hand-held ophthalmoscope, individual can independently take images of the fundus without the assistance of anyone else and obtain a prompt result established by researcher computation. The computation is useful for home consideration, remote pharmaceutical options, and self-examination.

(Gavin Robertson et al., 2011) Using a history characterized by blood glucose levels (BGLs), supper admission, and insulin infusions, Elman intermittent neural systems (ANNs) predicted BGLs. These presumptions originated from a worldview dominated by BGLs. Twenty-eight datasets were compiled using AIDA, a freeware numerical diabetes test system, and were structured using a single instance. Midnight of each day in 24 hours was optimal for making reliable predictions. Root mean square error was calculated in which system ensures the minimum error rate that ensures the system

reliability. BGL predictions as short as one hour out were tracked using a standard deviation of mmol/L. For the anticipated duration of BGL, which was up to 10 hours, an of (SD) mmol/L was tracked. In future wider variety of AIDA case scenarios, real patient data, and data identifying other factors that influence BGLs. To account for the ever-changing physiology associated with diabetes, studies will also be conducted on ANN standards based on continual repeating learning.

Developing a convolutional neural system is improvement-based and supported by a personal computer for diagnosing thyroid diseases using SPECT images (Liyong Ma et al.,2019). Graves' disease, Hashimoto's disease, and subacute thyroiditis are the three categories of illness considered here. Graves' disease is the most common of these three. The DenseNet engineering of the convolutional neural system is applied, but with some improvements, the preparation method is also enhanced. When doing so in DenseNet, adding the trainable weight parameters to each skip association causes a modification to the design. In addition, the preparation method is enhanced by increasing the learning rate with bloom fertilization calculation for system preparation, which results in an improved preparation approach. According to the findings of the clinical trials, the method proposed for the convolutional neural system is efficient for diagnosing thyroid infections using SPECT images, and has superior performance compared to other CNN systems.

Millions of people worldwide suffer from vision impairment due to diabetic retinopathy (DR) (Usman et al., 2023) because of the prevalence of diabetes. Avoiding blindness from untreated diabetes mellitus (DM) requires prompt diagnosis of DR. Color Fundus Photographs (CFPs) underwent extensive data pre-processing using

Principal Component Analysis (PCA) for feature extraction. PCA attempts to boost computing efficiency and model performance through dimension reduction and information preservation. The ML-FEC model was proposed to aid in the early diagnosis and treatment of DR and used convolutional neural networks (CNNs) that had already been trained. The experiment results were encouraging, with the ResNet50 architecture obtaining 93.67% accuracy, the SqueezeNet1 architecture achieving 91.94% accuracy, and the ResNet152 architecture achieving 94.40% accuracy. These findings prove that the suggested methodology can successfully identify and categorize DR lesions. To better diagnose and treat DR, this study fills a gap between data analytic methods and clinical applications.

Data mining, machine learning, and metaheuristic algorithms were utilized to develop a method for detecting type 2 diabetes (Al-Tawil et al., 2023). Effective management and preventative planning rely heavily on early detection. To improve detection methods, the cuttlefish algorithm, a bio-inspired metaheuristic algorithm, was proposed to identify crucial features during the preprocessing of medical data. The cuttlefish algorithm was used because of its usefulness in selecting features from medical datasets while remaining computationally efficient. To increase the precision of the detection process, the features that had the most effect on categorization outcomes. 70% of the data set was used as a training set for the cuttlefish and genetic algorithms. Logistic regression cost function and correlation-based feature selection were two criteria used to determine the optimal collection of features. The selected subset was then fed into various classifiers to boost detection precision.

ResNet50, a variant of ResNet, was used for feature extraction and classification on the ImageNet dataset (Abbood et al., 2022). With its many layers and residual blocks, ResNet50 facilitates more efficient training of deeper networks. The ResNet50 model requires around 3.8 billion calculations to solve. By deleting the Fully Connected (FC) layers and adding the Global Average Pooling layer, the Dropout layer, the Dense layer, and the SoftMax activation function, the authors adapted the ResNet50 architecture to the needs. With improved images and fine-tuned parameters gleaned from labeled data, the proprietary ResNet50-based model is trained to learn and detect important features associated with diabetic retinopathy. The model's performance was measured using several measures, including the accuracy rate (93.67%) and the hamming loss (a measure of how well the model generalizes). The updated ResNet50-based model is an efficient tool for detecting and categorizing DR-related lesions, and also applicable to clinical practice and large-scale DR screening programs.

Information is extracted from images of diabetic foot ulcers using a deep recurrent neural network (DRNN) architecture (Sathya Preiya and Kumar, 2023) to create a reliable approach for forecasting the progression of DFUs and facilitating timely and appropriate treatment. With a weight matrix comprising block matrices with diagonal sub-matrices, the DRNN architecture is well-suited for processing sequential data. The network's output is computed using forward propagation and activation functions, which mold the network's reaction to input data. During training, the weight update mechanism preserves the block diagonal structure to maximize speed and practicality. The major goal of the DRNN-based feature extraction technique is to capture and represent significant information from images of diabetic foot ulcers to improve the efficacy of classification tasks for predicting the development of DFUs.

2.3.7 Review of Classification Phase

S. Naz, H., & Ahuja, S. (2020) created a predictive tool for early diabetes identification utilizing machine learning techniques on the PIMA dataset. This addresses the ever-increasing incidence and catastrophic consequences of diabetes and compares Deep Learning classifiers, Decision Trees, Artificial Neural Networks, and Naive Bayes. With an accuracy rate of 98.07%, Deep Learning is the most promising method for developing an automated predictive tool to assist healthcare providers in detecting and preventing diabetes at an early stage.

Alhussan, A. A., et al. 2023 presented a fresh method for diabetes classification that combines an optimized random forest classifier with a new feature selection algorithm, dynamic Al-Biruni earth radius and dipper-throated optimization algorithm (DBERDTO). Diabetes is a common chronic condition, and this helps with the need for better diagnosis and prognosis. DBERDTO outperforms current optimization techniques and machine learning algorithms, reaching a remarkable 98.6% success in diabetes categorization. Statistical studies have shown that this method is superior.

Atteia, G. et al. 2022 used Bayesian optimization for tuning two custom CNN models for detecting diabetic maculopathy in retinal fundus and optical coherence tomography (OCT) images. This helps to develop a solution for the automation and speedy diagnosis of diabetes mellitus. The convolutional neural networks optimized by Bayesian methods outperformed pre-trained models, such as VGG16Net and AlexNet. The statistical analysis proved that the Bayesian optimization technique is efficient in deep learning model improvement towards better performance in detecting DM from retinal images.

For effectively diagnosing Type 2 Diabetes and refining the symptom data set, an Enhanced and Adaptive Genetic Algorithm-Multilayer Perceptron (EAGA-MLP) model was put forth by (Mishra S. et al. 2020). There are two aspects in which the EAGA-MLP works: proper disease identification and effective optimization of attributes for analysis in medical datasets. The EAGA-MLP model achieves a classification accuracy of 97.76% for identification of diabetic patients, with a 1.12-second runtime and also proves effective across seven other illness data sets and outperforms current algorithms that help health professionals diagnose diabetes more precisely.

(Shrinivasan L. et al. 2023) used the PIMA Indian diabetes dataset to provide a hybrid classification model for early diabetes prediction that combines decision tree algorithms with adaptive neuro-fuzzy inference systems (M-ANFIS). This improves diagnostic accuracy and reduces errors in diabetes onset prediction. Compared to conventional decision tree methods, the M-ANFIS model achieves better classification accuracy and is more efficient in detecting diabetes at an early stage.

(MuthuKumar and Jayakumar 2024) proposed a Deep Learning system for diabetes detection based on optimization, known as DeO-DiDet. IoT and machine learning were used to automate the process of diagnosing diabetes. An updated Marine Predator Optimization Algorithm handled feature selection, while a Gated Deep Recurrent Network handled detection. Seagull optimization improved classifier accuracy. With this, the model DeO-DiDet aimed to diagnose diabetes with velocity and precision while reducing the complexities of processing and time of calculation.

(Aliyu H. A. et al. 2024) proposed diabetic classification by applying a metaheuristic optimization strategy to address issues concerning dataset imbalance and classifier parameter tuning. Genetic algorithm and particle swarm optimization were deployed to optimize classifiers through data balancing. This is an effective way of improving the accuracy of diabetes classification. This was evidenced by the improvement in Average Precision Recall by 32.78% achieved when PSO-balanced data was used.

(Ashour et al. 2024), using feedforward and CNN against the Pima Indians diabetes database, identified cases of diabetes. The aim of this project was to increase diagnostic accuracy in healthcare by availing AI. The early and correct diagnosis of diabetes is the problem to be solved. The results demonstrate that the FNN model outperformed the prior techniques with an accuracy of 82%, while the CNN model reached 80.52%. When tested using binary classification for the diagnosis of diabetes, both models performed admirably.

To overcome the difficulty of making a precise remote diagnosis using electronic medical records (EMRs), (Yan F. et al. 2024) created a disease detection system for innovative healthcare. The system used fuzzy clustering, feature selection, and self-organizing fuzzy logic to enhance the quality of the data. The cleaned-up data is subsequently categorized using a machine-learning method that relies on eigenvalues. The system's efficacy in illness diagnosis is demonstrated by its evaluation utilizing datasets for cardiovascular disease and diabetes.

Abnoosian K. et al. 2023 suggested a new multi-classification system based on pipelines for diabetes prediction using the unbalanced Iraqi Patient Dataset. This

tackled issues like imbalanced datasets and limited labelled data by combining different pre-processing methods and machine learning models. This employed weighted ensembles based on AUC to improve performance. Results show high accuracy (0.9887), precision (0.9861), and AUC (0.999), outperforming other models and demonstrated the potential for accurate diabetes diagnosis and improved patient care in challenging data environments.

(Li, X. et al. 2023) suggested a hybrid feature selection method to improve diabetes diagnosis accuracy, focusing on the PIMA Indian diabetes dataset. This enhanced diagnosis accuracy, which is essential for efficient patient care, particularly amid the COVID-19 pandemic. The results demonstrated that the HR-K-means hybrid method outperforms traditional algorithms and prior studies regarding accuracy, reaching a maximum accuracy of 91.65%.

(Rehman A. et al. 2020) proposed a deep extreme learning machine model for diabetes prediction. Improved accuracy in diagnosing this common yet potentially fatal disease is very much sought after. This DELM approach achieved high reliability with minimal error, thus outperforming previous artificial neural network models. As evidence of its effectiveness at diabetes prediction and achieved an accuracy rate of 92.8% when used with 70% training and 30% test/validation data.

(Le T. M. et al. 2020) proposed a wrapper-based feature selection technique and an early diabetes prediction technique based on the Multilayer Perceptron (MLP) model, Improved with APGWO: Adaptive Particle Greywolf Optimization. This improved the accuracy of classifications due to optimized feature extraction of relevant attributes

from the medical records. From all the models developed based on machine learning and was found that the APGWO-MLP model was the most efficient model, with an F1 score of 98% and a maximum efficient accuracy of 97%.

(MOUSA A. et al. 2023) suggested comparing three deep learning models for diabetes diagnosis using the Pima Indian Diabetes Database: The three algorithms chosen are as follows: Long Short-Term Memory (LSTM), Random Forest (RF), and Convolutional Neural Network (CNN). As the model has comprehensive features and benefits, this Pima Indian Diabetes Database is applied to assess model performance. Among all the models that have been trained and tested, LSTM yielded the best results, followed by RF and CNNs.

(Shao et al. 2024) developed Optuna LightGBM, merging the LightGBM algorithm with automated hyperparameter optimization using Optuna tools to predict diabetes. This enhanced the model accuracy and precision while dealing with imbalanced datasets issues. Results indicated an improvement in accuracy from 97.07% to 97.11% and precision from 97.17% to 98.99%. This identifies the model's effectiveness in attaining predictive performance.

(Akhtar et al. 2024) in diabetes prediction, optimization techniques for feed-forward box neural networks (FFNNs) were applied, which involved the oppositional whale optimization algorithm (OWOA). This tackled the local optimal problem, which affects global optimal solutions for complex medical data and employs a dataset of diabetic patients from the Pima Indian community to compare optimization strategies. This demonstrates how AI and metaheuristic techniques can influence how medical diagnoses are made and manifest relevant treatment forms.

(S. M. Ganie et al. 2023) used boosting techniques to predict diabetes with the Pima database to create a system for its early detection through machine learning. After preprocessing and assessing the efficiency of five algorithms, gradient boosting achieved an accuracy level of 92.85%, which is higher than other studies. The research illustrates the effectiveness of this approach in predicting diabetes as well as other diseases with adequate markers. The model's performance was evaluated using several metrics, including precision, recall, and receiver operating characteristic (ROC) curves.

To mitigate the problem of late diabetes mellitus (DM) diagnosis in areas with a deficit of healthcare professionals, (Rufo, D. D. et al. 2021) used the data from Zewditu Memorial Hospital, located in Addis Ababa, Ethiopia. This employed a Light Gradient Boosting Machine (LightGBM) on the understanding that this does not require many resources and minimal computational cost. In the ZMHDD dataset, the sensitivity of the model was at 99.9%, that of specificity was at 96.3%, while the Area Under the Curve (AUC) was at 98.1%, which was better than SVM, KNN, Bagging, NB, RF, XGBoost.

For diabetes diagnosis, (Thompson, C., & Higgins, O. 2024) have suggested a technique called artificial neural network-particle swarm optimization (ANNPSO) with the aim of enhancing diagnostic accuracy. PSO was primarily employed to optimize the weights of ANN to improve its ability to successfully discover hidden patterns in patient data. This combined approach improved diagnostic tests' sensitivity, specificity, and accuracy. This result outperformed all the other traditional methods by a large margin, achieving 94.15% accuracy.

(Ghabousian R. et al. 2024) suggested a method for diabetes diagnosis that combines a fuzzy system with the Particle Swarm Optimization (PSO) algorithm. The fuzzy system binarized the PSO to improve performance and is employed to optimize feature selection. This improves the accuracy of diabetes diagnostic classifications in the Pima Indian dataset. The outcome is a classification precision of 95.47 percent, which is faster and more accurate than competing methods.

To control blood sugar levels in people with type 1 diabetes, (Dagher, K. E., and Haggege, J. 2024) suggested an optimizing adaptive auto-tuned PID controller that uses PSO and GWO. While stabilizing glucose levels between 60 and 120 mg/dl, these meta-heuristic approaches optimize control parameters for accurate insulin delivery, reducing tracking errors to a minimum. Improvements in control efficiency and patient safety are evident from the results, which provide a 4-12% faster stabilization time than conventional approaches. Additionally, there is a considerable decrease in overshoot and oscillation.

(Ulutas H. et al. 2024) utilized a hybrid strategy for hyperparameter optimization based on PSO-GWO and ensemble learning methods like boosting, bagging, voting, and stacking to enhance diabetes diagnosis. Combining many classifiers into a single stacking ensemble model yielded the best results (98.10% accuracy), followed by random forest. Because of these findings, better and more dependable diabetic diagnostic systems are possible.

(Pradhan G. et al. 2024), the proposed forest architecture jointly selects features with a Binary Multi-Neighborhood Artificial Bee Colony (BMNABC) and classifies them using an ensemble of Random Forests. This would maximize the depreciation of several diabetic mellitus (DM) identical attributes from five DM datasets and early detection into sections. This achieved an accuracy of up to 97.28% and a sensitivity of up to 99.95%, outperforming other approaches while being flexible and efficient in different clinical settings.

An Oppositional Whale Optimization Algorithm (OWOA)-tuned Feed-Forward Neural Network (FNN) for diabetes prediction was presented by (Chatterjee R. et al. 2024). OWOA is employed in situations when conventional optimization methods are unable to converge rapidly enough or get trapped in local minima. By fortifying the exploration and exploitation phases, OWOA increases accuracy and convergence speed. The findings show that OWOA outperforms WOA, ABC, and PSO as the best algorithm for identifying diabetes.

The ISSA technique was created by (Hassan, G. S. et al. 2023) by using the Salp Swarm Algorithm (SSA) to fill in the missing values in the Pima Indian Diabetes Disease (PIDD) dataset. Naïve Bayesian Classifier (NBC), SVM, and KNN score higher in classification, and the results demonstrate that ISSA works better than traditional imputation methods such as eliminating samples or replacing non-existent or non-zero values, the mean, or random numbers.

The classical PSO intentionally performs while searching for local convergence. So, according to A. U. Rehman et al. (2020, the convergence speed of the classical PSO is

improved for the multi-cluster jumping problem. The swarm population is separated to search for the global optimum. The jumping techniques avoid local minimum trapping to relocate the particles in a random position. The evaluation results show that the convergence speed of the new clustering approaches has improved significantly for high-dimensional datasets.

Local outlier factor (LOF) and swarm intelligence are combined to develop a hybrid optimization technique for outlier detection in high-dimensional streaming data (A. Karale et al., 2020). This outlier detection approach helps reduce memory utilization and time complexity while processing the streaming data.

Karayılan et al. (2017 utilized the backpropagation neural network (BNN) to predict heart disease. In this, 12 hidden layers are created to automatically analyse the Cleveland heart disease dataset variables to predict heart disease. This BNN model attained a maximum 95.5% accuracy rate in predicting heart disease.

Ebrahimi-Ghahnavieh et al. (2017 combined the Convolution Neural network (CNN) model features with features; CNN performs feature learning, and Long Short-Term Memory is used to perform Alzheimer's classification. In this study, the performance of nine different CNN models is verified to learn the pathological features of Alzheimer's. The CNN-based Long Short Term Memory model attained a maximum accuracy rate of 90.62%.

Amarbayasgalan T. et al. (2019) introduced a deep learning model that predicts cardiovascular disease risk. The deep model utilizes the auto encoder to estimate risk using the reconstruction error data. The entire dataset on cardiovascular disease is

partitioned into a training set and a testing set for the purpose of prediction. The experimental outcomes show that the auto-encoder-based method attained a maximum 86.337% ROC score.

Zak M. and Krzyżak A. (2020) introduced the VGG 16 net, ResNet 50, and Inception V3 Net, which are utilized to segment and classify the pathological regions of lung disease from chest x-ray images. In this, the non-segmented and segmented chest X-ray images are trained to verify the disease classification performance of the U-net-based deep model. The VGGnet and Inception V3 Net attained 95% classification accuracy for segmented images.

Fatima et al. (2020) analyse various machine learning techniques, data mining approaches, and deep learning systems to predict breast cancer. The primary goal of this research is to examine existing breast cancer datasets in order to increase recognition accuracy. In addition, this analysis helps beginners classify breast cancer by using various machine learning techniques.

Zhang et al. (2020) introduced the advanced ensemble classifier with a linear discriminate analysis (LDA) approach to recognize breast cancer. Gene expression data is processed here using the auto encoder and LDA approaches to extract features. The derived features are investigated using an ensemble deep-learning approach that classifies breast cancer with up to 98.27% accuracy.

Byra et al., 2021 Predicting breast cancer from an ultrasound image using Siamese convolutional neural networks. This process uses deep convolutional neural network-based learning to predict cancer with a maximum recognition rate. Here, ImageNet database images are utilized in L1 regularized logistic regression model processes to

extract the various patterns, and double transfer learning is applied to classify the breast mass. Cancer is recognized by the identified features compared with the morphological features.

Prostate cancer detection from histopathology pictures using multiscale decision consolidation and data augmentation, D. Karimi et al., 2020. The deep learning technique with three patch size-related approaches extracts the various features. The derived features are examined using a convolution network with a logistic regression model. This process recognizes the disease with minimum time consumption and a high accuracy of 92%.

Utilizing a fuzzy deep-learning method, S. Park et al. (2016) predicted intra- and inter-fractional variances in the analysis of lung cancers. The introduced method minimizes computation time while predicting lung tumours. This process uses the breathing clustering algorithm and fuzzy deep learning techniques to predict lung tumour variations with 83.17% accuracy. In addition, the introduced method attains a minimum computation time of 1.54ms while predicting the inter- and intra-fractional variations.

Liang et al. (2019 apply reiterative learning approaches to perform biomedical image segmentation. This process uses gastric cancer images processed by a reiterative learning framework. The introduced method identifies the affected region by computing the manual annotation data. This method is achieved by extracting the various image characteristics, and the weak annotations are derived. The derived information was used to improve the accuracy of the biomedical image segmentation by up to 91.09%.

Nasir Khan et al. (2019) created the mammogram classification by fusing multi-view features and convolutional neural networks. Here, four-view mammogram images are used to predict cancer with high accuracy and minimum computation time. This process uses the CAD tool to extract image features and cancer information effectively. The system utilizes the mini-MIAS database and curated breast imaging subset images to evaluate the system's effectiveness.

Du et al. (2019) apply Bayesian deep multi-view learning approaches to reconstruct the human brain's activities related to perceived images. This process computes the statistical relationship between brain views according to the specific generators. The derived visual image generations are processed using the deep learning approach, which classifies the human visuals. Then a Bayesian inference approach is applied to recognize the visual images from the fMRI dataset with maximum accuracy.

Vivek Kumar et al. (2021) identified the morbidity from clinical notes using ensemble deep-learning approaches. Clinical notes for a patient are processed with techniques including word embedding, bag-of-word processing, and feature selection. Deep learning uses the resulting features as input to detect patterns with efficient learning functions. The defined system was evaluated using the n2c2 natural language processing dataset.

Md. Rishad Ahmed et al. (2018 created a dementia disease diagnosis system from neuroimaging using deep learning. The primary objective of this research is to examine the work of numerous researchers who have looked at the diagnostic process for

Alzheimer's dementia. Dementia can be detected precisely using deep learning methods and the associated learning functions.

David Ahmedt-Aristizabal et al. (2021) investigate the risk rate of schizophrenia in children based on the EEG response using a deep learning approach. The EEG is analysed, and abnormal features are extracted and processed by a recurrent deep convolution network, classifying the children's health risk rate. The defined system uses children aged 9 to 12 years, and the system's effectiveness is determined using cross-validation.

Ali et al. (2020) identified the motor units using a deep-learning approach from musculoskeletal ultrasound images. This process examines the images and estimates the signals transferred into the spatiotemporal images by applying deep neural networks. This framework predicts the pipelines based on the spatiotemporal characteristics effectively.

Chen et al. (2020) applied fuzzy convolution with a hidden Markov approach for identifying and locating the vertebrae. This process is used to resolve the difficulties in feature extraction and sequence modelling in the long sequence of 3-D image analysis. The problem is overcome by extracting the image components by applying the encoding and decoding processes. Then various contextual information is retrieved that helps locate the vertebrae. The discussed system ensures 87.97% accuracy in the identification rate in the MICCAI 2014 dataset.

D. Pham et al., 2019: Detecting Parkinson's disease from short-time series using deep learning fuzzy recurrence plots. The long-short-term memory neural network is utilized to examine the large volume of features necessary to classify the disease patterns successfully.

Aliasing artifacts in ultrasound colour images were reduced using deep learning techniques by H. Nahas et al., 2020. The CFI artifacts are initially examined using a convolution network that segments the affected region. Then aliasing is identified by applying U-net architecture, which minimizes the difference up to 2.51%.

Tuncer et al. (2020) implemented the multiclass-pathologic voice classification process by applying a multi-centre and multi-threshold-based ternary pattern-based voice algorithm (MCMTTP). Here, different multi-level features are derived via the sample-related discretization technique and neighbourhood. The neighbourhood component analysis method is then applied to the generated features for further investigation. The final step involves using MCMTTP to analyse the retrieved features to diagnose the pathological condition based on the vocal pattern.

In 2020, S. Kaur et al. developed an AI-powered medical diagnostic system. This study investigates the various researcher’s work collected from January 2009 to December 2019 by classifying more than 1–5 articles. The author employs nearly efficient machine learning algorithms like deep learning and fuzzy logic. Machine learning methods are applied to the data for the most precise diagnosis of organ failure (kidney, liver, prostate gland, brain, and heart).

Ezzat et al. (2020) introduced a gravitational search-optimized deep-learning methodology for diagnosing COVID-19 disease. The optimized network processes the chest X-ray images to diagnose COVID-19. Here, the DenseNet121 architecture is applied to detect the affected rate of COVID-19 using a social ski driver. This process ensures up to 94% accuracy on the test set.

Ramesh S. et al., 2020 recognized paddy leaf disease with the help of Jaya-optimized deep neural networks. Different plant leaves are collected and processed by pre-processing techniques to eliminate unwanted noise and details. Then clustering algorithm is applied to segment the disease-affected portion. Further, different features are extracted and processed using a defined algorithm. The algorithm detects the leaf disease with 98.9% accuracy on various diseases and normal leaves.

De La Torre et al. 2020 created the diabetic retinopathy disease grading system using interpretable deep-learning approaches. The defined classifier analyses the retinal images to recognize the disease severity level. Then score values are provided to the features, and the pixel-wise score propagation model is applied to neurons for effectively recognizing diabetic retinopathy.

In 2020, Fouad et al. looked into how using an AI system could improve the efficiency of patient assistance programs by analysing data collected from Internet of Things-based health monitoring devices. An optimized-deep belief network based on the golden section processes the collected IoT data. The new method can detect 99.87% accurate variations in patient health situations.

AlZubi et al. 2020 presented the tubu-optimized sequence modular network for predicting Parkinson's brain disorders from IoT-based collected data. The created system reduces the false classification accuracy and improves the disease prediction rate with maximum speed.

Agrawal S. et al. 2020 predict breast cancer using the optimized deep neural learning technique. The learning process uses features such as location, tumour size, and types to improve cancer recognition accuracy. The system's effectiveness was determined using the Wisconsin breast cancer dataset, and the system diagnosed cancer with minimum time and complexity.

T. Ibuki et al. 2020 created the flocking control for multiple rigid bodies using optimized techniques. The rigid bodies are studied using the three flocking laws of cohesion, alignment, and separation. In addition to this, alignment rules are applied to resolve distributed optimization problems and collision avoidance.

S. Mohan et al., 2019 developed a hybrid machine learning technique-based heart disease prediction model. Here, IoT devices are used to record the patient health details, which are processed with the help of a hybrid random forest linear model approach that predict the disease with 88.7% accuracy.

Huang et al., 2019 maximize the feed-forward network classification efficiency by combining gravitational search, fuzzy rules, and particle swarm algorithm. This process uses the mesothelioma and chronic kidney disease dataset. Then the features are extracted and processed by the feed-forward network. Network weights and bias

values are optimized during this process to eliminate the maximum deviations. Here, the optimized parameter values are selected according to the fuzzy rules.

Developing a swarm intelligent optimized regressive neural network by Lozito et al. 2019 to improve the classification process. Intelligent techniques are utilized to resolve the high explorative and starling optimization problems. The optimized techniques train the networks to achieve the minimum error rate.

2.3.8 Review of Flock Optimization Algorithm

Flock optimization algorithm (Xie et al., 2020) is defined as the particle swarm algorithm that works according to the bird social behavior. The algorithm comes under the swarm intelligence techniques that utilized in the decentralized systems for addressing the optimization problems. The basic idea behind the flock optimization technique (Laudani, A. et al., 2014) is particle communication and collaboration procedure to find the optimal solution in the search space. The optimization algorithm has different phases such as swarm behavior and exploration-exploitation. In the search space, each bird in the flock is defined as solution which used to address the optimization issues. During the searching process, flock adjust their position according to their movement and experiences. For every movement, the gathered information is frequently shares between the flocks to identify the best solution. The searching process balances the exploitation and exploration while predicting the optimal solution that ensures the convergence. The flock searching process identifies the neighboring information to predict the particles and the identified solutions used to determine the best solution. For every selection of solutions, the flock position is updated and again

searching process is initated until to reach the convergence. This food searching process of flock optimization algorithm manages the adaptability while addressing the optimization problems Bahaidarah, M. et al., 2024.

H. S. Baruah et al., 2020 utilized a new particle swarm optimization (PSO) to select the relevant feature sets from eight different datasets. The habit of bird flocks influences the PSO. Mutual information is used to assign a score using the feature selection method. The evaluation results show that the optimization method boosts classification accuracy across the board.

Xu et al., 2022 predict soil compression coefficient (SCC) according to the behaviour of the crow flow optimized feed-forward learning process. The system uses the flight length, flock size, and awareness probability values to examine the SCC with 76.19% accuracy.

Gedam, A. N. et al., 2024 applying flock based optimized deep convolution neural network for identifying lung nodules. This approach uses the fishing characteristics to identify the best region from image. The selected region features processed by the deep convolution network which recognize the lung nodules with maximum recognition rate and convergence speed.

Bansal, J. C. et al., 2023 introduced drone flock optimization algorithm with principal component analysis for addressing the multi-objective problems. The main intention of this work is to eliminate the collision in the industry applications by reducing the feature dimension using the principal component analysis. The optimization algorithm

used to control the industrial parameters that solve the objective problems with maximum recognition accuracy.

Chou F. I. et al., 2021 classifying heart disease using particle optimization with XGboosting techniques with controllability of fractional order approach. The main intention of this work is to enhance the robustness, reliable and flexible solutions while predicting heart disease using the behavior of flocks. The gathered information is processed using XGBoost classifier and the performance of the system is improved by updating the optimization approach which minimize the error rate and improve the overall efficiency.

Zhang et al., 2018 predator-prey particle swarm optimized algorithm for detecting Alzheimer's disease. The brain images are processed, stationary wavelet entropy-related features are extracted, and the particle swarm optimized technique is applied to select the optimized feature. This process improves overall disease recognition with up to 92.69% accuracy.

Introducing the optimized Crow algorithm-based Parkinson's disease detection system in Gupta et al., 2018. An optimized classifier processes the collected information by resolving the high-dimensional data processing issues. The system's efficiency was measured against 20 benchmark datasets.

In 2020, Cherian et al. presented a hybrid lion-optimized neural network for cardiac illness prediction. Several statistical characteristics of the heart are retrieved from the collected data. The derived features are processed by combining the particle swarm

algorithm lion algorithm. This optimization technique minimizes the deviations while classifying the heart disease. The system classifies the heart disease effectively compared to the traditional algorithm.

Kaur et al., 2020 predict Parkinson's disease by applying hyper parameter-optimized deep-learning neural networks. Here, various hyper parameters are utilized for training the deep networks, and the process is optimized using a grid search algorithm. The mined features are treated by a layer of networks and functions that recognize the disease with 91.69% accuracy on speech samples.

2.4 Previous Work for Benchmark Analysis

The current methods for detecting diabetes and the corresponding implementation procedures are discussed in this chapter. The paper draws on the recommendations and findings of "Convolutional Recurrent Neural Networks (CRNN) for Glucose Prediction" by Kezhi Li et al. (2019) to inform its methodology. This study employs the convolution neural network to foretell glucose concentration. One of the key reasons for diabetes discovered early to prevent risk factors is the glucose level. Glucose homeostasis is the root cause of diabetes. Pancreatic insulin secretion is regulated by -cell and -cell hormone transfer, which maintains normal blood sugar levels. If people are infected due to Type-1 diabetic disease, the $\beta - cell$ have cooperated, which affects insulin production. This causes hypoglycemia (blood glucose concentration < 70 mg/ dL) and hyperglycemia (blood concentration > 180 mg/dL). Then insulin therapy is required to manage the blood glucose level, and a continuous glucose monitoring system is developed. According to the importance of the blood glucose level, the research uses the CRNN system to compare the introduced

Flock Optimization Assimilated Deep Learning Model (FOADLM) based diabetic prediction model. This chapter deals with the CRNN glucose level monitoring and diabetic prediction process, and the working process of FOADLM is explained.

2.5 Diabetic Prediction model

This section clearly explains the working process of the Convolutional Recurrent Neural Networks (CRNN) Diabetic prediction and glucose monitoring process. The detailed data collection process and respective data analysis are explained to identify the research gap in the future diabetic prediction process.

2.5.1 Data Acquisition and Settings

In CRNN systems, two Clinical and in-silico datasets are utilized to analyze the glucose monitoring process. The in-silico dataset has ten adult information (T1DM) collected with the help of the UVA/Padova T1D simulator. The simulator analyzes the glucose level that the Food Drug Administration grants. The author uses the modified simulator version during the data collection to support the variability characteristics. The variability is measured on insulin absorption, meal composition, insulin variability, absorption, and carbohydrate estimation. Additionally, simple physical exercise is considered while analyzing the patient's glucose level.

Clinical data is captured in London Marys Hospital-Imperial College Healthcare Trust. T1D subjects are continuously examined in the clinical trust, and the insulin bolus level is identified. The patients are continuously observed for six months, and the insulin level is collected. Therefore, the dataset consists of insulin, glucose, time

stamps, and meal levels. The observed clinical dataset also consists of the exercise time stamp, insulin boluses, and meal information. The collected information is fed into the computation model to identify the glucose level, and the model improves the Continuous Glucose Monitoring (CGM) efficiency. In addition, the computation model uses different inputs, such as alcohol consumption, stress, and exercise, to improve computational efficiency.

The In-silico dataset utilizes the UVA/Padova T1D simulator to collect patient health information. The validated and robust framework is used to create the simulated cases during the data collection. The simulator uses different set-ups to collect the information because each patient has different insulin and meal levels. Ten unique adult cases are generated for each case, which has been checked for three meals/day. The insulin values vary daily (1 to 5) and are recorded with and without a meal. Of the collected data, 60% is used for training, and the remaining data is utilized for testing.

The clinical data is gathered from T1DM subjects by conducting 6-month clinical trials. The patients are continuously monitored using the Dexcom G4 Platinum CGM sensor. The sensor gathers the information every five minutes, and the collected details are utilized to analyze the glucose level. The CGM sensors are inserted into the patient's body, and manufacturing instruction is followed while gathering the information. In addition, diabetic patients transfer the exercise, insulin, and meal level to analyze the patient's health condition. However, the sensor and simulators successfully record the patient health information; few records are missing due to manual and machine problems.

2.6 Existing Methods Utilized in the Diabetic Prediction models

The CRNN system uses preprocessing feature extraction and time-series prediction phases. These phases are more useful for improving the diabetic prediction rate. The gathered information consists of inconsistent information affecting the entire prediction model. As a result, the preprocessing procedure is used to cleanse the dataset of its inconsistencies and extra characteristics. After the data has been cleaned and organized, feature extraction is used to obtain useful statistics. The author extracts the features using Convolutional Neural Networks (CNN). Time series predictions are made using Long Short-Term Memory Neural Networks (LSTM), and the overall accuracy of the prediction is enhanced using a signal converter. Figure 2.8 depicts the general structure of the CRNN used in the diabetic prediction procedure.

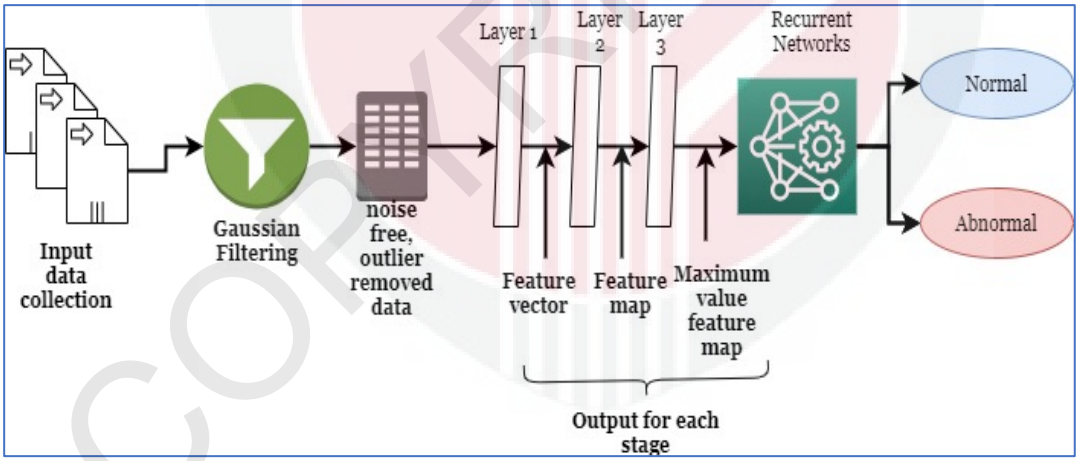


Figure 2.8: Operational Structure of the CRNN

The system uses the glycemic data collected from the CGM sensors, insulin, and carbohydrate information. In addition, alcohol, exercise, and stress-related information are utilized as input. The input is processed using the preprocessing technique that removes the noise and produces the cleaned time-aligned glycemic, insulin, and

carbohydrates data as the output. The preprocessed output is fed as the input to the Convolution Neural Networks (CNN), and the output is fed into the Recurrent Neural Networks (RNN) input. The RNN output helps predict the glucose level, and the hidden states are continuously updated to minimize the deviation between actual and predicted values. The model is continuously trained during this process to improve the overall prediction rate. The system efficiency is evaluated using 30- and 60 min prediction processes because the glucose prediction software runs easily. Then each phase involved in this process is explained in the following section.

2.6.1 Data Preprocessing

This step aims to remove outliers and extraneous information from the data and formatted appropriately for processing by a neural network. The author uses normalization and time stamp alignment to improve data quality. In addition, interpolation, filtering, outlier detection, and extrapolation are performed to eliminate irrelevant information from the dataset. The clinical data consist of several outliers and missing values, continuously influencing the prediction efficiency. Therefore, the outlier identification process is applied to eliminate irrelevant information. The author uses the one-dimensional Gaussian kernel filtering approach to eliminate the missing and inconsistent data from the dataset. The filtering technique only applies to the clinical dataset because the in-silicon information is already preprocessed and cleaned. The clinical data is analyzed using the gaussian filter, which takes two hours to change the data structure. During this process, 24 window size is utilized as the sliding window that completely removes the clinical data's noise. After removing noise from the diabetic information, the transferred input is fed into the next feature extraction phase.

2.6.2 Feature Extraction

Feature extraction is deriving the statistical details from the noise-removed information. The extracted features are representations or characteristics of the diabetic disease. The author uses Multilayer Convolution Neural Networks (CNN) to extract meaningful features. The time-series input data is transformed into features useful in diagnosing diabetes using the CNN method. The network uses the convolution process to compute the feature value, and the convolution process is computed using equation (2.1).

$$Z[m] = \sum_{i=-l}^{l} x[i].\delta[m-i] \tag{2.1}$$

In equation (2.1), the input signal is denoted as x, the kernel value is represented as δ, and the convolution of the input signal is defined as $Z[m]$; here, m is the index of the respective convolution value. The convolution network uses the different layers that predict the output values for input x. The first layer uses the 24 sizes of the sliding window with x' Length and a kernel size of 8. Then the network layer utilized dimensions are defined in Table 2.5.

Table 2.5: Convolution Neural Networks-Layer Dimensions

Layer	Output Dimensions	Number of Parameters
Convolution layer $(B * S * C)$: B-Batch; S-Step and C-Channels		
$1 * 4$ convolution	128(1) *24*8	104
Max-pooling, size =2	128(1)*12*8	-
$1 * 4$ convolution	128(1)*12*16	528
Max-pooling, size =2	128(1)*6*16	-
$1 * 4$ convolution	128(1)*6*32	2080
Max-pooling, size =2	128(1)*3*32	-
Recurrent Layer $(B * C)$		
LSTM	128(1)*64	24832
Dense layer $(B * U)$; U-Units		
Dense	128(1)*256	16640
Dense	128(1)*32	8224
Dense	128(1)*1	33

According to Table 2.5, the inputs are processed in every layer, and the output is predicted. The input is fed into the sliding window settings; the window may be non-overlapped or overlapped. The window condition is determined according to the convolution network computations and size. The neural model learns the features automatically by using the weights and interconnections. Then the process of the convolution neural network features is illustrated in Figure 2.9

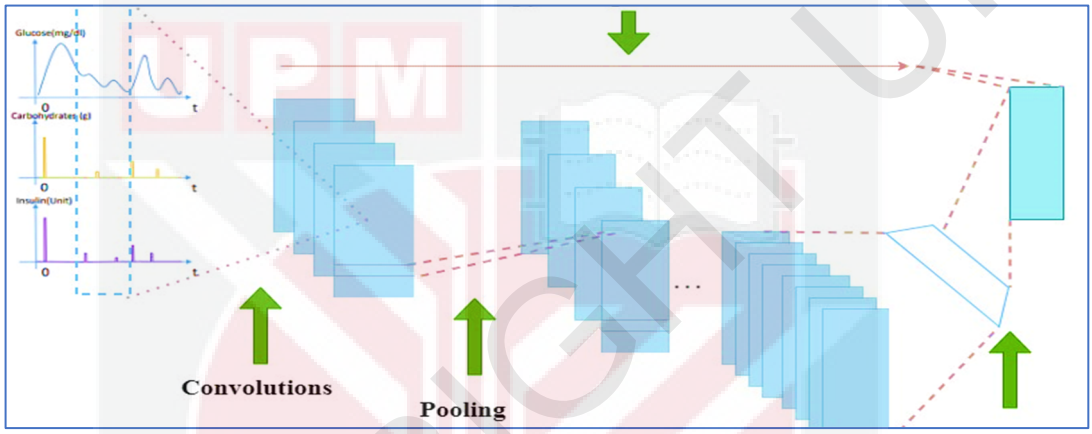


Figure 2.9: Graphical Structure of the Convolution Neural Networks for feature extraction of the noise-removed clinical data

Convolution neural networks, as shown in Figure 2.9, are depicted along with the structure and its function. The neural network model consists of multiple layers, including a convolutional layer, a max-pooling layer, and a fully connected layer. After a feature map is formed from the convolution layer's outputs, the data's dimensionality is reduced using a max-pooling layer. A pooling layer is introduced between the convolution layers to reduce the data dimension and boost the total data analysis and prediction rate. The insertion of the pooling layer reduces the overfitting issue and maximizes the diabetic prediction rate. Considering the accepted size is $L1 * D1$, this has been down-sampled by different parameters such as stride and Spatial

Extent F. Then the max-pooling process produces the output vector Y with the size of $L2 * D2$.

$$L2 = \frac{(L1-F)}{S+1} \tag{2.2}$$

$$D2 = D1;\ Y_i = \max(y_i^*) \tag{2.3}$$

In equation (2.3), y_i^* is denoted as the down-sampled output value, feature map is represented as Y_i and the maximum value of the features is denoted as the max () operator. The convolution network trains the data using stochastic gradient descent and a backpropagation algorithm. The training reduces the difficulties while extracting the features from the time-series input data. The network continuously uses the updated random weight values during the computation, minimizing the error rate and classification problem. The weight value is denoted as w_i and network bias value is b_i. The last output from the convolution layer is fed into the recurrent layer, which makes the diabetes prediction.

2.6.3 Classification

The author uses the Long Short-Term Memory (LSTM) recurrent network to predict the diabetic disease time-series data. The LSTM comprises 64 cells with input, forget, and output gates. Every gate has a set of neurons and functions to perform the output computation. The LSTM network has predictive and time series function that identifies the sequential and time series data. In addition, each cell has a memory function that stores every processing input and output value that helps to improve the overall output identification efficiency. The LSTM network is linked to the multi-dimensional time series output via a single hidden layer. Each layer comprises 64 cells, and a dropout

layer follows the RNN network. during the analyse, need to reduce the number of neurons used for prediction and training by using a dropout layer. The dropout layer mitigates the overfitting problems, and generalization is improved. Figure 2.10 then depicts the basic layout of the LSTM network.

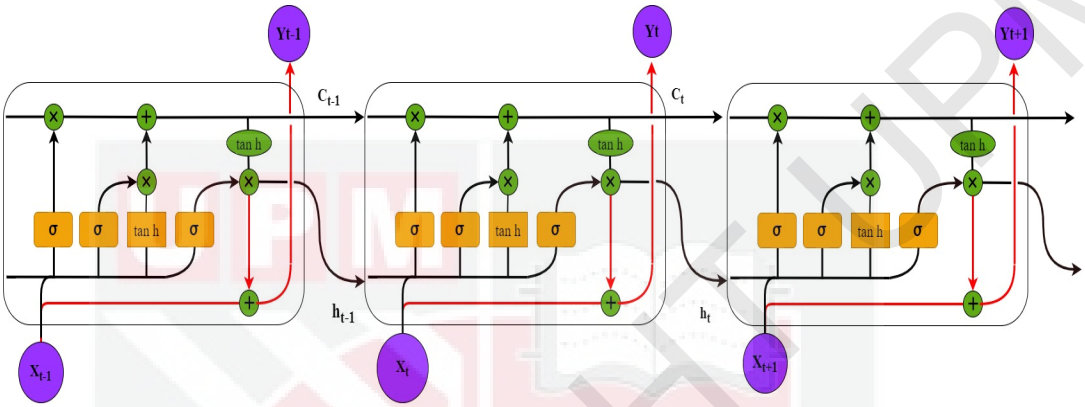


Figure 2.10: Long Short-Term Memory Neural Network Structure

The LSTM has transformation and recovery steps to compute the output value. During the training process network uses the backpropagation learning between the current $x(t)$ and future $x(t+6)$. The way of backpropagation learning utilized in this step is the transformation step. Then, the sliding window has multi-dimensional time series data containing meals, blood glucose values, and insulin. After training the data, the output changed between $x(t)$ and $x(t+6)$. Then the blood glucose value is estimated using a 2..4.

$$x(t+6) = x(t) + \Delta x(t) \tag{2.4}$$

According to equation (2.4), the recovery step is computed to predict diabetic disease. During the output estimation, LSTM computes the conditional probability value for the data sequence defined in equation (2.5).

$$Condition\ Probability = p(y_1, y_2, \ldots\ldots y_{T'} | x_1, x_2, \ldots\ldots x_T). \quad (2.5)$$

In equation (2.5), $x_1, x_2, \ldots\ldots x_T$ is defined as the input and $y_1, y_2, \ldots\ldots y_T$ is represented as the output sequence. The sequence has T and T' in size; the input vector is defined as x_t, output vector is defined as y_t and memory cell vector is defined as c_t. The network has parameters for every computation, such as weight W, U, and bias b. These parameters are applied on three gates, such as forget gate f_t, input gate i_t and output gate o_t. Then the computation for every gate is defined in the following equations (2.6 to 2.12).

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (2.6)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2.7)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (2.8)$$

$$g_t = \sigma_g\left(W_g x_t + U_g h_{t-1} + b_g\right) \quad (2.9)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_t(g_t) \quad (2.10)$$

$$h_t = o_t \circ \sigma_t(c_t) \quad (2.11)$$

$$y_t = o_t \circ \sigma_t(c_t) + x_t \quad (2.12)$$

In the above equations (2.6 to 2.12), the sigmoid function is denoted as σ_g, the hyperbolic tangent function is defined as σ_t and the entry-wise product is denoted as $\circ$. h_t is defined as the hidden internal state fed into the next step. The y_t is defined as the network output value and x_t is denoted as the input. Blood glucose levels are determined by comparing the computed value y_t with the baseline glucose value, which helps improve the overall diabetic prognosis. At last, the recurrent network processes the input with two hidden layers (256 and 32 neurons) to predict the single neurons. The final output is estimated using equation (2.13).

$$Z_i = act\ (\sum_{i=1}^{N} Y_i w_i + b_i) \tag{2.13}$$

In equation (2.13), the multi-dimensional output is defined as Z_i, the activation function is represented as $act()$, fully connected network has different parameters such as w_i and b_i. The linear function is used to calculate the output value by CRNN in this work. Equation (2.14), used to estimate the sigmoid function,

$$act(a) = a \tag{2.14}$$

Furthermore, the network employs the gradient descent optimization process to reduce the discrepancy between the true and anticipated values. The optimizer boosts the overall prediction rate with the help of the RMSprop. Table 2.6 then defines the Kezhi Li system's algorithmic steps.

Table 2.6: Algorithm for CRNN Diabetic Prediction Process

Algorithm for Existing System
An existing technique utilizing the following steps: pre-processing, feature extraction using CNN, time series prediction using LSTM, and a signal converter.
Algorithm:
Initialize: input I ($i_1, i_2, i_3, \ldots . i_n$); features F ($f_1, f_2, \ldots . . f_n$), weight w, bias b
(pre-processing)
Step 1: convoluting I with $G_\sigma(x, y)$ // Gaussian function $G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}$
The output of step 1: filtered input (fI)
(Feature extraction)
Step 2: Transfer fI into multilayer convolutions to get the feature vector F.
// defined convolution as z[m]=$\sum_{i=-l}^{l} x[i].\delta[m-i]$,
$x[i]$ Input, kernel - δ $(size\ 8)$, convolution result is z with index m, in the first layer. //
Step 3: Transfer the z[m] into the next convolution layer with the pooling function to obtain the feature map (y^*) which is obtained after performing down-sampling in that convolution layer.
Step 4: Then maximum feature map (Y) is obtained by examining (y^*).
Step 5: The final Y-predicted convolution layer fed into the recurrent layer
(Prediction)
// here LSTM network is trained by using respective inputs and outputs, which helps to predict the incoming inputs pattern//
//The network uses the input vector, output vector, and memory cells to analyse input data//
Step 6: Fed Y into the input, forget, and output gates.
Step 7: Get Y_{out} value from the output gate.
Step 7: Update (w and b)
Step 8: compute the overall out using the sigmoid function (Z_i)
Step 9: Predict the patterns as abnormal or normal.

According to the above algorithm steps, the glucose level is predicted continuously. Here, CRNN successfully utilizes the convolution neural networks, recurrent neural model, and linear activation function to predict the glucose level. In addition, each procedure is applied with the same precision to boost the automatic system's total performance. The CRNN approach able to handle both temporal and spatial dependencies data features. The combination of these features helps to handle the complex pattern while analysing the large volume of data. The CRNN recurrent layer derives the exact features from the input that helps to identify the exact spatial patterns from the time-series and medical images. The hierarchical feature extraction process contributes the model to improve the overall diabetic prediction rate. The recurrent layer in the CRNN has memory unit remember entire sequential information that used to monitor and predict the patient health information in real-time scenario. In addition, the integration of recurrent and convolution networks identifies the complex relationship between data which helps to address the overfitting issues successfully.

As a result, this study uses the CRNN system as its foundation to develop a reliable method for detecting diabetes. The CRNN system analysis identifies the research gap described in Table 2.7. In addition to the CRNN system, the introduced Flock Optimization Assimilated Deep Learning Model (FOADLM) is compared with different existing models such as Convolution Recurrent Neural Networks (CRNN) (Kezhi li et al. 2020), Deep Neural Networks (DNN) (Nadesh et al., 2020), improved Particle swarm optimization optimizing Long short-term memory network (IPSO-LSTM) (W. Wang et al., 2020), gravitational search optimized deep learning (GSODL) (Ezzat et al., 2021), tubu optimized sequence modular network (TOSMN) (AlZubi et al., 2020), Adaptive Neuro-Fuzzy Inference System (ANFIS) (F. Haque et al., 2021), Ensemble Learning Approach (ELA) (N. L. Fitriyani et al., 2019) and

DiabDeep (H. Yin et al. 2021) methods. A brief discussion of each existing method is described in the below section.

2.7 Comparative Methods

This research uses almost seven hybridization methods (models) to compare the introduced Flock Optimization Assimilated Deep Learning Model (FOADLM). The short description of each algorithm is explained with the working structure, process, advantages, and disadvantages. The explanations are more helpful in understanding the introduced FOADLM efficiency.

2.7.1 Deep Neural Network (DNN)

The first comparison method is obtained from Nadesh et al., 2020 "Type 2: diabetes mellitus prediction using deep neural networks classifier." The author utilizes the Deep Neural Network (DNN) to predict type 2 diabetic disease. The system intends to improve the disease prediction rate. During the analysis, n hidden layers are utilized to examine the diabetic data. The working process of DNN in the Type-2 diabetic disease detection process is illustrated in Figure 2.11

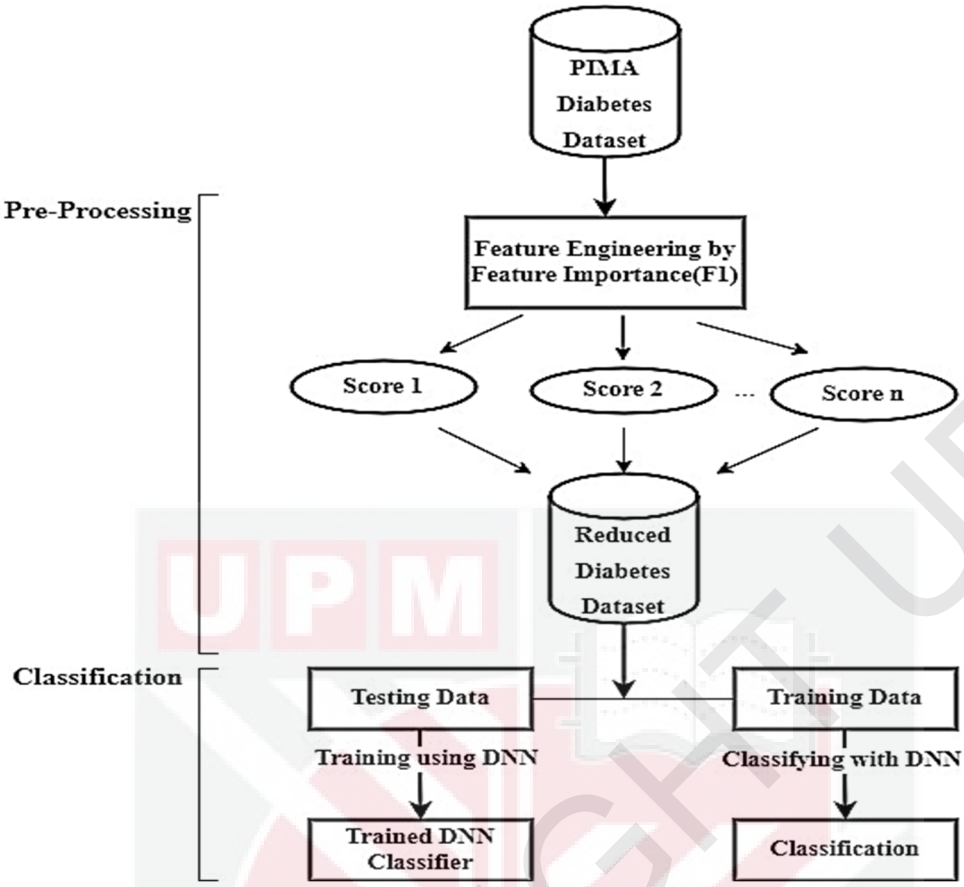


Figure 2.11: DNN-based Type-2 Diabetic Detection Process

Nadesh utilizes the PIMA Indian diabetic dataset information to identify diabetic disease. The collected data is processed with the help of the Feature Importance Framework (FIF) to predict the feature subset. Before, the author used pre-processing to remove irrelevant, noisy, and inconsistent information. The noise removal process removes ambiguous data and redundancy. The noise-removed data is analysed to extract the features to predict diabetic disease with maximum recognition accuracy.

Table 2.7: Research Gap Identification from Kezhi Li System

Author & Year	Title	Method	Performance	Limitation	Datasets
Kezhi li et al. &2019	Convolutional Recurrent Neural Network for Glucose prediction	Modified CRNN, LSTM	Best result in the simulated dataset than the real-time dataset	The modifications are common regardless of the data stream and quantity. This results in non-adapting computations, requiring multiple repeated validations (high prediction horizons). Performance Measure: RMSE, Accuracy	Real-time datasets of the patient and simulated dataset
Kezhi li et al. &2020	GluNet: A Deep Learning Framework for Accurate Glucose Forecasting	GluNet: Personalized deep neural network with probabilistic distribution	The RMSE of the Proposed model got good results compared to SVR (Support Vector Regression), LVX (Latent variable with exogenous input), NNPG (Neural network for predicting glucose), and ARX (auto regression with exogenous input). The above 4 are the methods used for comparison. Performance Metrics: Computation Cost and RMSE	High-quality training data is required to retain prolonged accuracy results. The validation instances increase the time lag.	In-Silico, Clinical datasets

The introduced FIF approach helps to identify the diabetic-relevant features from the dataset. The FIF is also named the Extremely Randomized Tree, which analyses the feature list features and selects the most relevant features. The approach explores each feature by computing the entropy value, information gain, and high score level. These characteristics are considered to foretell the optimal characteristics. The tree model is constructed using a feature selection technique that reduces overfitting problems and increases the accuracy with which diabetes may be predicted. Type-2 diabetes was then classified using the capabilities of deep neural networks. In this case, the ReLU activation function is used to derive the result. According to the output value, normal and abnormal features are classified effectively. The DNN system has several advantages such as easily predict diabetic disease from the PIMA dataset, reduced overfitting and redundancy issues in DNN based diabetic prediction process and enhance the overall recognition accuracy. However, the system has few disadvantages such as requires additional optimization techniques to minimize the output deviation, optimized feature selection techniques require to improve the overall recognition accuracy and difficulties in processing the large volume of diabetic data.

2.7.2 Improved Particle Swarm Optimization Optimizing Long Short-Term Memory Network (IPSO-LSTM)

Next, W. Wang et al.'s 2020 "Blood Glucose Prediction with VMD and LSTM Optimized by Improved Particle Swarm Optimization" work compares the introduced Flock optimization model. The collected blood glucose information is processed with the help of a variation method decomposition that analyses the blood glucose level at different frequencies. The LSTM neural network is applied to classify diabetic disease. The VMD approach identifies the optimal solution for the variational problem by

transferring from construction into the solution domain. During the analysis, the author uses the Hilbert Transform, Wiener Filter, and Frequency mixing to improve the overall diabetic disease detection efficiency. After cleaning up the time-series data, LSTM is used to make predictions about people with diabetes. Parameters of the network, including the weight and bias value, are optimized with the help of Particle Swarm Optimization. The updating process is done according to the particle position and velocity. The optimization-based network fine-tuning procedure minimizes the classification problem and improves recognition accuracy. Then the overall working process of W. Wang et al. 2020 diabetic prediction procedure illustrated in Figure 2.12.

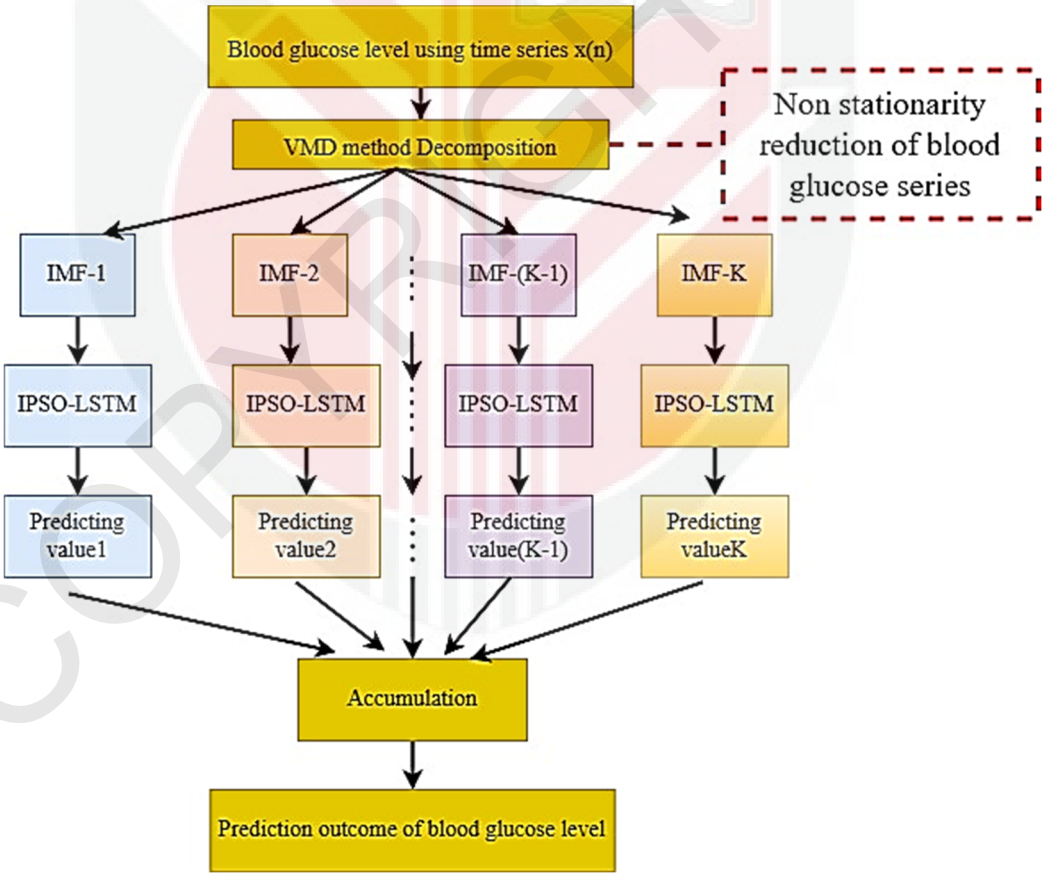


Figure 2.12: IPSO-LSTM-based Diabetic Prediction model Structure

According to Figure 2.12, diabetic data is collected with the help of a CGM sensor that gathers the patient's blood glucose value. The collected information is normalized to [0,1] to simplify the prediction process. The information is split in two: 500 records are used for testing, while 1000 are used for training. Then VMD is applied to decompose the data into non-stationary and non-linear values according to the frequency value. The LSTM is applied to the frequency band to predict the diabetic value. Here, the PSO approach is applied to optimize the LSTM performance, minimizing the computation difficulties and deviation between the output values. The IPSO-LSTM has various advantages such as reducing the variation problems while analysing the time series of data, improve the diabetic detection rate by normalizing the input values and minimizing the difficulties in feature and parameter learning. Even though, the system has various disadvantages such as maintaining physical characteristics such as liver function, age, and gender while analysing diabetic disease. The system only concentrates on the short period of data to predict diabetic disease and consumes high computational difficulties while analysing large volumes of diabetic information.

2.7.3 Tubu Optimized Sequence Modular Network (TOSMN)

AlZubi et al. 2020, analysed and created a system to identify Parkinson's disease from wearable IoT data using the optimized neural network. Parkinson's disease affects people's movement, expression, and speech gradually. The author aims to predict Parkinson's disease with maximum accuracy and a false detection rate. Internet of Things (IoT) data can be fed into the Deep Brain Simulator (DBS). The gathered information is processed using the TOSMN approach that predicts every change in

brain function. The modular neural network divides the input into different modules, and each module output is estimated separately. The output is combined to get the final output value which helps to minimize the prediction delay and increases the computation speed. The overall working process of the tubu-based modular neural network process is illustrated in Figure 2.13

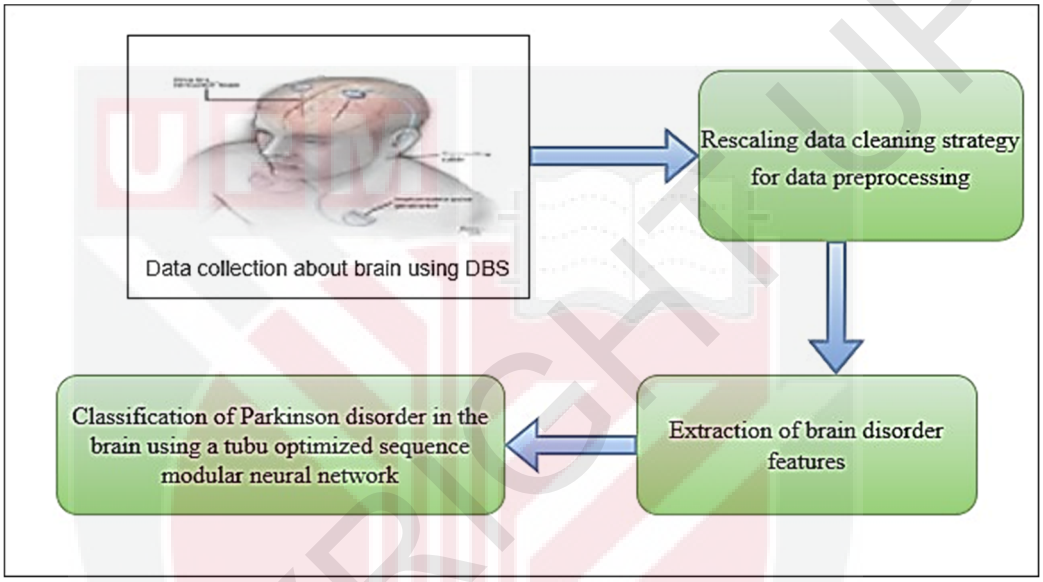


Figure 2.13: Tubu Optimized Sequence Modular Network-based Parkinson's disease detection process

Figure 2.13 illustrates the working process of the TOSMN approach, which uses different stages such as data collection, pre-processing, feature extraction, and Parkinson's disease classification. The author uses the rescaling data cleaning strategies to explore every data in the dataset. The rescaling process changes the dimensionality and normalizes the data to reduce the computation difficulties. Then different statistical features are extracted from the brain data. A modular neural network that can detect even the slightest alterations in brain activity is used to process the data obtained. The tubu optimization method is applied during the data analysis to optimize the neural network function by updating the network weights and bias value.

Thus the TOSMN approach maximizes the overall Parkinson's disease detection rate and reduces the computation difficulties. The TOSMN has several advantages such as system can process real-time data with maximum recognition accuracy, minimize the false classification rate while analysing the large volume of brain data and resolves the classification maximization problems. Although the system requires high computation time while analysing the candidate list in the search space. Increases the processing time for each module, maximizing the computation difficulties.

2.7.4 Gravitational Search Optimized Deep Learning (GSODL)

Ezzat et al., 2021 introduced Gravitational Search Optimized Deep Learning (GSODL) approach to recognize the Covid-19 disease. The author uses DensNet to analyse the collected Covid-19 related health information. The DenseNet uses the Convolution Neural Network that uses different layers, such as the convolution layer, pooling layer, and fully connected layer. These layers investigate the input X-ray images using the hyperparameter that identifies the disease with maximum recognition accuracy. During this process, the gravitational search algorithm is applied to update the network parameter. The selected parameters are utilized in the convolution network training model. The network froze the parameters during the data analysis, improving the overall image classification efficiency. This process uses the Binary Covid19 dataset information during the efficiency analysis, and the system attains 95% accuracy. The system uses the Inception v3 network to extract the network features, and hyperparameters are quantified while comparing the results. From the comparison, the author introduced attains 95% of accuracy. Then the overall architecture process is illustrated in Figure 2.14

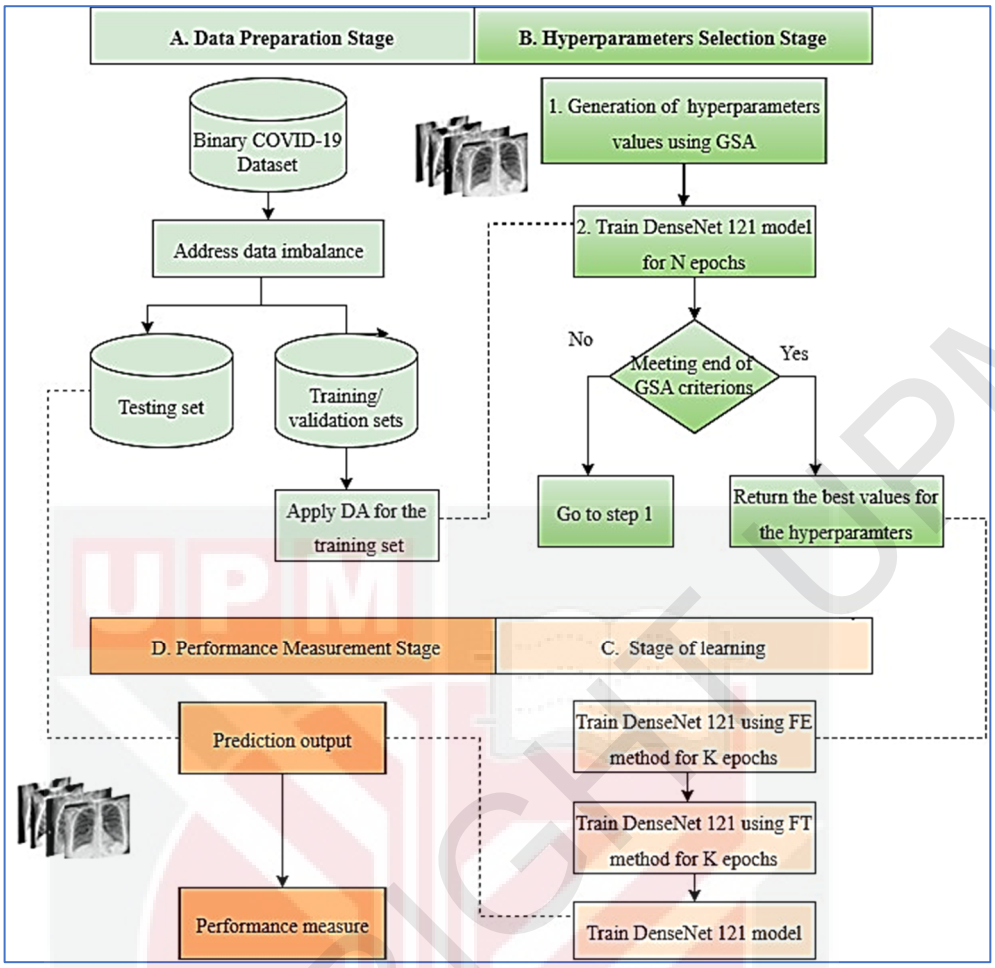


Figure 2.14: GSODL approach based Disease Prediction Process

Figure 2.14 illustrates the GSODL approach based covid-19 disease detection process. Here, the author detects the disease in two stages: data analysis and hyperparameter selection process. The data is first gathered and analysed to address the disparity in data issues, which then contributes to greater illness detection precision. After cleaning up the data and was separated into sets for training and testing. The divided data is processed using the densenet that identifies the disease information. The gravitational search algorithm selects the hyperparameter that minimizes the overall deviation between the computed outputs. The GSODL effectively reduce the data imbalance issues and improves the overall data analysis accuracy. The algorithm effectively utilizes the data augmentation and optimization process that minimizes deviation

between the outputs. The densenet approach uses data splitting techniques that lead to creating the replication of data, which maximizes the computation time and maximum computation and network memory overhead.

2.7.5 Adaptive Neuro-Fuzzy Inference System (ANFIS)

F. Haque et al., 2021 introduced Adaptive Neuro-Fuzzy Inference System (ANFIS) to classify Diabetic sensorimotor polyneuropathy. The author intends to improve patient health and reduce the mortality rate. This work uses the Michigan Neuropathy Screening Instrumentation to collect patient input. The algorithm has five layers with directional nodes and links. The algorithm uses the forward and backward passes to compute the output value. A Hybrid Learning algorithm is applied to improve the data analysing efficiency during the analysis. The effective utilization of the hybrid learning process recognizes the patient information into different categories such as severe, moderated, mild and absent. During the analysis, the Epidemiology of Diabetic Intervention and Complication dataset information is utilized to analyse the patient's health details. The clinical trials using the neuro-fuzzy inference system minimize the error rate and maximize the diabetic prediction rate. The network also explores the data that identifies the gait from the tibialis anterior, vastus lateralis, and gastrocnemius medialis. The overall working process of the ANFIS system is described in Figure 2.15

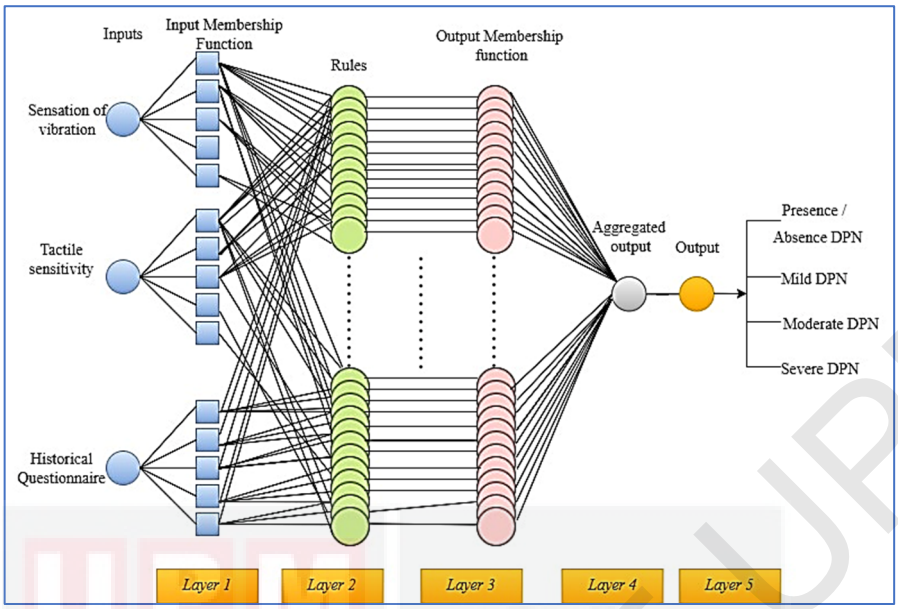


Figure 2.15: ANFIS-based DPN prediction process

Figure 2.15 illustrates the working process of the DPN prediction process using ANFIS. The neural model inputs the sensation vibration, tactile sensitivity, and historical questionnaires. Different layers and outputs process the inputs are aggregated. During the analysis, fuzzy logic is applied to predict the exact outputs, such as the absence/presence of DPN, mild, moderated, and severe DPN. The ANFIS process has various advantages such as withstand the real-time data process and attains maximum stratification and identification accuracy. The system could predict the severity of the diabetic disease with minimum computation complexity. Even though the system requires the patient's previous conditions to analyse the future health status and difficult to manage the large volume of data and consumes high computation complexity. The system requires an optimization technique to analyse the high-risk factors related to information.

2.7.6 Ensemble Learning Approach (ELA)

N. L. Fitriyani et al., 2019 recommended the Ensemble Learning approach to predict diabetic disease and hypertension. The author intends to develop an effective diabetic prediction model to reduce the difficulties involved in noise-data analysis. The isolation Forest method processes the gathered diabetic data by excluding the outlier. In addition, the minority oversampling technique is applied to investigate the data distribution. The overall working process of the ELA-based diabetic prediction process is illustrated in Figure 2.16. The ensemble learning process reduces the weak feature involvement in diabetic disease prediction. The created system uses the four datasets to analyse the system's efficiency. The ensemble learning-based developed system reduces the risk factor and improves the overall disease detection efficiency.

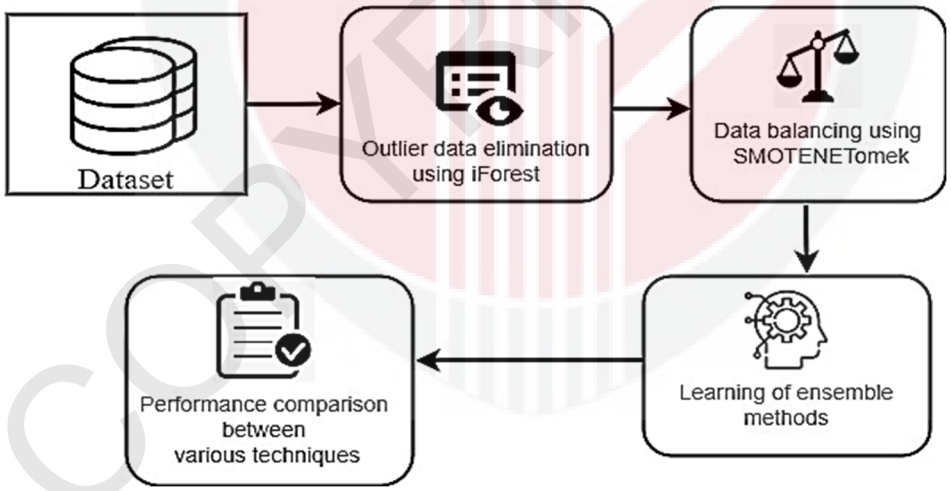


Figure 2.16: ELA-based Diabetic Prediction model

Figure 2.16 illustrates the ELA approach-based diabetic prediction process. The author uses the Chronic Kidney Disease (CKD), prehypertension, hypertension, and type 2 datasets. The dataset contains 403 patient information, and 19 attributes are utilized to

compute the output value. The collected information is processed using the SMOTETomek to resolve the imbalanced issues. The algorithm explores every piece of information to maximize the minority class instances, reducing the noise from the image. The ELA approach improves the data classification efficiency by reducing the weaker features. Then overcomes the data imbalances issue and outlier issues. even though the system required optimized parameters to create the generalized and robust model. The data sampling process must be improved to handle the different criteria-related data.

2.7.7 DiabDeep

H. Yin et al., 2021 introduced the DiabDeep approach to identify and detect a diabetic disease from sensor data. The author uses neural and medical sensor networks to predict the changes in people's health. The sensor devices on the human body continuously record the information transferred to the health centre. The recorded information was analysed using the neural network to process the features. The network uses layers, functions, and learning processes to predict diabetic features. This process uses the recurrent layer in the Long-Short Term Memory network to classify. The gradient learning and magnitude pruning approach is applied during the analysis to update the network parameters. The overall working process of the DiabEeep approach is illustrated in Figure 2.17

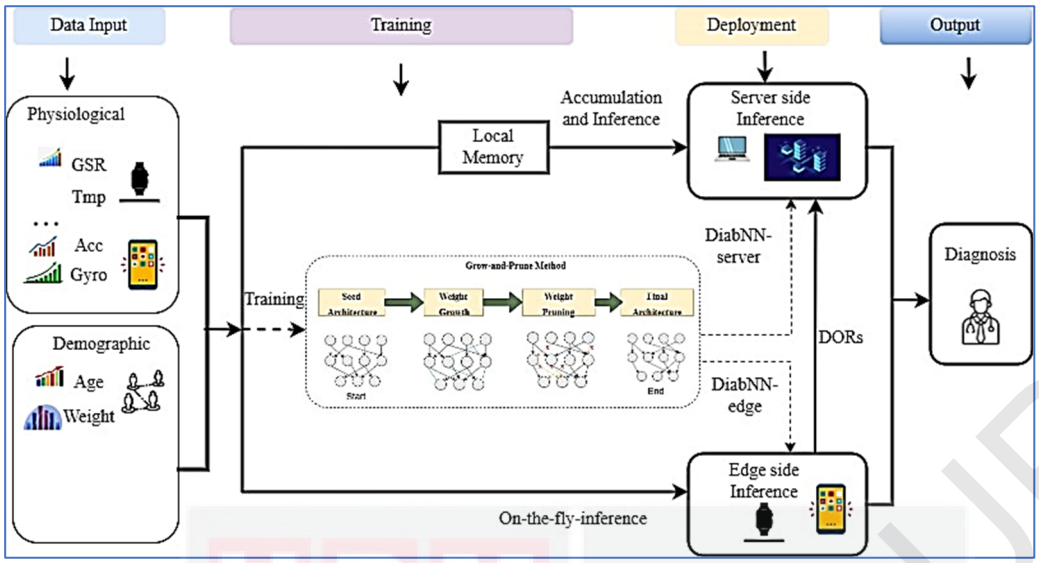


Figure 2.17: DiabDeep Based Diabetic Prediction Process

Figure 2.17 illustrates the working process of the DiabDeep approach-based diabetic detection process. This work uses physiological information, and demographic details are input into the network. The collected information was trained using a long short-term memory network that adjusts the network parameter's weight and bias to improve the overall detection accuracy. Then the testing features are processed, and the computed output is compared with the testing output. The created system efficiency is evaluated using 52 participant information, and the system recognizes the disease with up to 96.3% accuracy. Even though the system works, this method effectively has a few advantages and disadvantages, which are listed below. This process effectively manages the system's robustness and flexibility and improves the overall data prediction accuracy. However, the system should concentrate on the feature extraction process to meet the scalability factor while analysing the large volume of data.

2.7.8 Convolution Neural Networks-Long Short-Term Memory Networks (CNN-LSTM)

Swapna et al. introduced Convolution Neural Networks-Long Short-Term Memory Networks (CNN-LSTM) to detect a diabetic disease from heart signals and health data. The patients are continuously monitored to obtain the ECG signal's Heart Rate Variability (HRV). The collected heart information is investigated using the CNN network, which derives features from the ECG signal. The features are further analysed using the LSTM network that predicts the abnormality in health conditions. The created system was evaluated using a 5-fold cross-validation approach that recognizes the diabetic disease with 95.1% accuracy. The overall working process of CNN-LSTM based diabetic prediction process is illustrated in Figure 2.18.

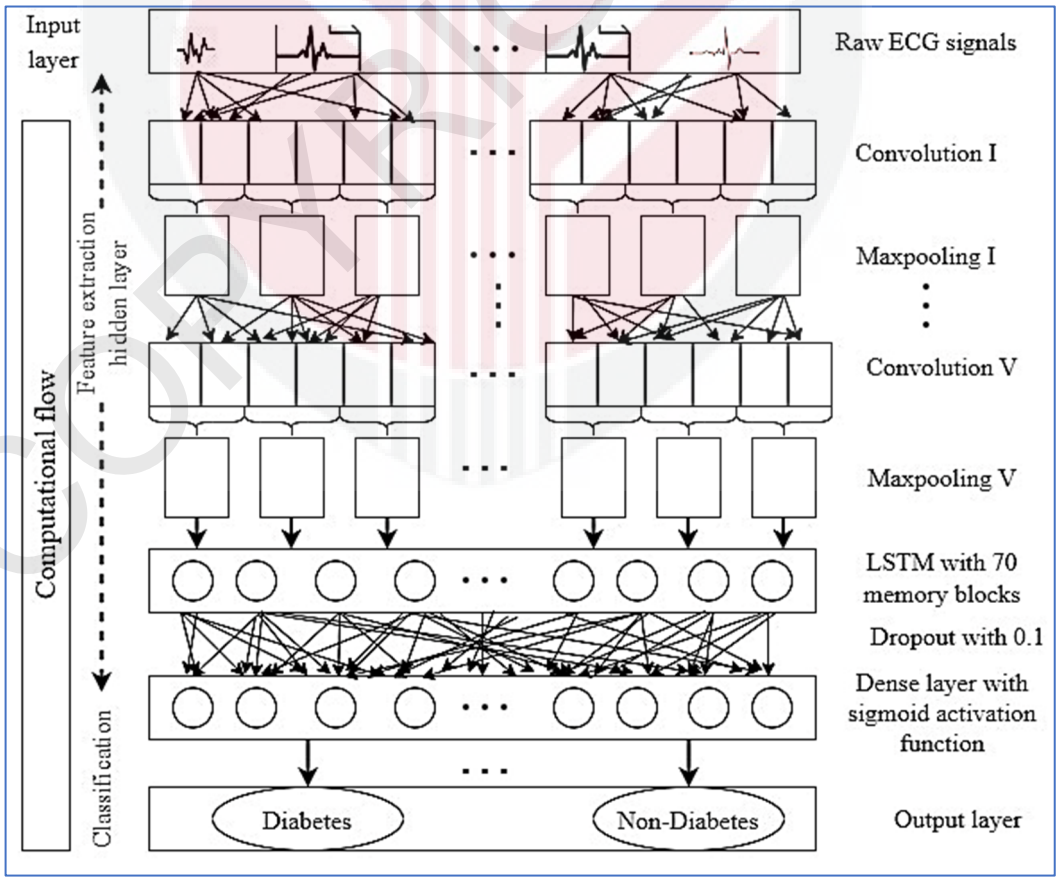


Figure 2.18: CNN-LSTM-based diabetic prediction model

Figure 2.18 illustrates the CNN-LSTM-based diabetic prediction model architecture. The network consists of several layers input, convolution layer, max pooling layer, LSTM with memory blocks, dropout layer, dense layer with sigmoid activation function, and output layer. The raw ECG signal is the network's input and is processed by the layer above. Each layer performs its function using activation functions. The convolution layer uses the kernel and stride functions to generate the feature map. A max-pooling layer is used to explore the feature map, decreasing the feature dimensionality and increasing recognition accuracy. Every processing input is stored in the LSTM memory block, minimizing computation difficulties in future data analysis. Finally, the dropout layer is applied to reduce the irrelevant feature involvement. The sigmoid activation function is then used to predict the output value. The CNN-LSTM reduces the computation difficulties while processing a large volume of data. Stores every processed input in memory that makes system computation easier and maintaining data processing robustness and flexibility. However, the system difficult to maintain high accuracy while analysing large volumes of data.

Similar works in the area of machine learning and deep learning have assessed different models and optimizations for improved predictive accuracy, most notably time-series analysis, classification problems and optimization problems. The machine by Kezhi Li et al. (2020) Convolution Recurrent Neural Networks (CRNN) Several Convolutional layers are used for feature extraction while Recurrent layers (RNNs) are used for sequential data processing, this allows for the effective usage of temporal data. The Deep Neural Network (DNN) introduced by Nadesh et al (2020) has gained reputation for its capability to learn highly complex relationship and is mostly used to conduct pattern recognition task. Wang W. et al (2020) proposed Improved Particle

Swarm Optimization-Long Short Term Memory based on improved Long Short Term Memory, successfully combines particle swarm optimization PSO with Long Short Term Memory for better forecasting application. The Gravitational Search Optimized Deep Learning (GSODL) by Ezzat et al. (2021) uses gravitational search methods in neural network training as parameter optimization to enhance efficiency of training and accuracy. The TOSMN machine by AlZubi et al (2020) improves the sequence modular neural network approach by integrating optimization techniques to undermine challenges faced in training modules separately. Adaptive Artificial neural network, Fkuroroet et al, (2005), All Consultants ANFIS: Fatura et al (2021) undermines the weaknesses of neural networks assisted fuzzy systems in data uncertainty management. Enhanced 'Ensemble Learning Approach' – ELA, Fitriyani et al (2019) believes possible to combine various learning algorithms to get better results. Finally, DiabDeep H. Yin et al (2021) is about deep learning methods for predicting diabetes such as other studies which show domain deep learning methods can work well. For the purpose of this study, these methods were selected to act as a benchmark as demonstrates to the present day evolution in machine learning and optimization that can be utilized in delivering different outputs in terms of framework, optimization strategies and application areas and also make reliable benchmark models because of the proven success in application to resolving intricate issues that are comparable to the ones addressed in this research analysis.

2.8 Summary

This chapter provides context for the study by examining the history of diabetes and its associated risk factors, symptoms, and consequences. To get a full detail of the

issue, performs a thorough literature analysis, looking at many sources from many angles. Authors' thoughts and insights and computational methods, methodologies, and approaches to the problems are also examined. Research problems, or specific questions or challenges, are identified and framed based on these ideas in this chapter. After examining the current literature, carefully create these issues to fill in the blanks and investigate under-explored areas. Research into these issues directs efforts toward developing novel computational approaches to existing difficulties. This chapter offers a solid comprehension of the setting by investigating the history of the disease, its contributing causes, and the current state of knowledge. The literature review deepens this comprehension by integrating many points of view, methods, and approaches. This broad comprehension directs the investigation of computational methods and tools for tackling the recognized difficulties.

CHAPTER 3

DEEP LEARNING MODEL BASED ON FLOCK OPTIMIZATION ALGORITHM

3.1 Introduction

This section discusses the Sampling Flock Optimization Assimilated Deep Learning Model-based diabetic prediction model. This proposed model's main idea is to deeply predict the diseases with the highest recognition accuracy using the sampling flock optimization deep learning model by analysing the drawbacks of different models and concrete as a standard fusion model with highly evaluated performance. The problem definitions were coined into two sections: non-sequential data due to scarcity and hard-level assessment due to data constraints. The issues in the above-discussed methods concern data scarcity (availability) and its quality in one measure. Contrarily, the computations and correlation operations generate high complexity and convergence in the prediction process. Therefore, the data sequence and assessment are expected to be precise in determining the prediction accuracy. Machine learning requires multiple training instances without intact optimization or pre-processing methods. However, the optimization level and the learning input must be balanced without convergence in a hybrid prediction process. This prevents unnecessary data augmentation, sequence variations, and prolonged convergence in handling heterogeneous data. Therefore, this study aims to improve the overall diabetic detection process.

3.2 Flock Optimization Assimilated Deep Learning Model

The existing system analysis has the health data scarcity and un-synchronization problem which affects the diabetic prediction rate. Availability is less due to fewer wearables and its operation time. The placement of the sensors causes this problem, due to which the computation complexity increases. The computations are based on different input patterns and rely on hybrid optimization systems. Therefore, the continuity problem and periodic data updates are considered a lag. Un-synchronized data streams are common due to the different operating hours of the sensors. Both issues result in non-sequential (missing) data that remains insufficient for precise accuracy. The validations and modelling is limited to the data's quantity and density; hence, the all-time precise output is less expected. According to the research gap analysis, the proposed methodologies are suggested, which is described in Table 3.1.

Table 3.1: Suggestion of Proposed Methodology

Title	Methodology	Advantages	Limitations	Proposed Suggestions/ Modifications
Convolutional Recurrent Neural Network for Glucose Prediction	Modified recurrent layer based Convolutional neural network	Less RMSE	The modifications are common regardless of the data stream and quantity. This results in non-adapting computations, requiring multiple repeated validations (high prediction horizons).	Sampled flock confines the validation instances depending on sequential and discrete data.
GluNet: A Deep Learning Framework for Accurate Glucose Forecasting	GluNet: Personalized deep neural network with probabilistic distribution	High forecasting accuracy, less complexity.	High-quality training data is required to retain prolonged accuracy results. The validation instances increase the time lag.	Training, testing, and validation is classified independently for the adaptable data streams identified. The variations are identified based on features and data association, reducing unnecessary validations.

Table 3.1 suggests the proposed diabetic prediction model, and the respective architecture is illustrated in Figure 3.1. The working process demonstrated that the system has data collection, pre-processing, feature extraction, selection, and classification process. Gaussian Filtering is utilized during the analysis to remove the noise from the data. Then the features and the data association are identified in the classification process for initializing independent flocks. The flocks generate differential solutions until the end-of sequence and non-replications. The identified sequences are trained independently using the deep learning process. Both attribute and non-attribute datasets are used for training the learning system. Depending on this process effective diabetic disease classification system is created to overcome the problem definition. The Flock Optimization Assimilated Deep Learning Model has been developed to solve the diabetes disease classification problem with respect to data imbalance and assimilation by employing the combination of optimization and deep learning paradigm. One of the issues that are usually experienced in medical datasets is class imbalance where the number of samples in two or more classes (for example diabetic and non-diabetic) differ widely which results in incorrect diagnoses. To counteract this, the model adopts the flock optimization techniques (Laudani et al. 2014) to improve the learning by changing the weights in the network, so that classes that are not well represented during the training receive enough attention. Furthermore, there is the assimilation problem: how to combine different types of features and data from different domains. This is resolved with the use of selection and transformation methods oriented toward the optimization process. This determines that the most important features useful for diabetic classification are learned while those that are unnecessary or contain noise are eliminated. As the class imbalance problem is solved and feature assimilation is enhanced, the Flock Optimization Assimilated Deep

Learning Model improves the accuracy, robustness, and generalization capability of diabetic disease classification model.

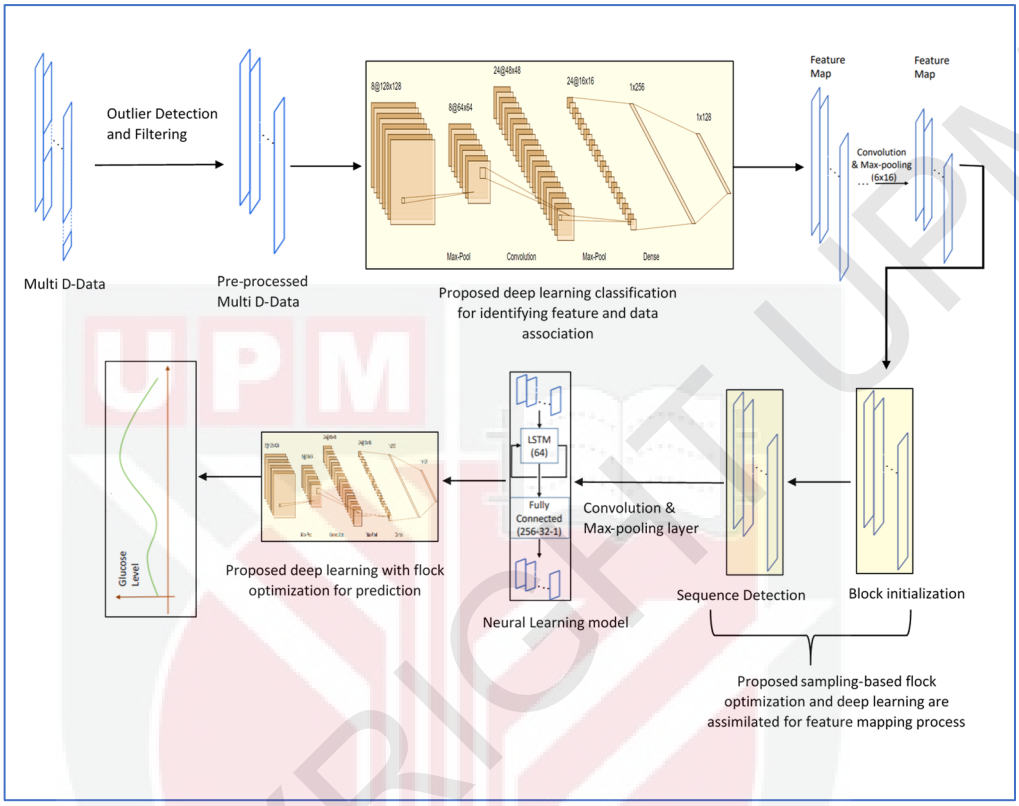


Figure 3.1: Flock Optimization Deep Modelling based Diabetic Prediction model

Figure 3.1 illustrates the flock optimization deep modelling-based diabetic predictions system. The multi-dimensional dataset collected is preprocessed using the Gaussian Filtering approach, eliminating irrelevant data. Then the sampling process is applied to normalize and improve the data representation. After the dataset has been prepared, a deep learning model extracts the features of interest. A feature map is created in the convolution layer based on the feature. Then the generated feature map is normalized using the sampling optimization model that predicts the feature association and solves the assimilation issue. The output is predicted by applying the Long Short Term Memory and fully connected layer to the reduced features. The function of the neural

model is optimized using the flock optimization (Tongur, V., et al 2019) method after the output values have been computed. This process minimizes the classification problem and data assimilation issues and improves the overall data classification accuracy. Even though, the FOADLM approach effectively detects the diabetic data, trade-off exists between reaching high accuracy on the model and maintaining generalization of the model. In the future, improvements may be made by using data generation methods, optimizing computation without compromising on the results, or incorporation of interpretability to increase the model's reliability.

3.2.1 Phases Involved in the Proposed Methodology

As discussed above, the proposed methodology has several steps listed below.

- Data Collection
- Preprocessing
- Feature Extraction
- Classification

3.2.1.1 Data collection

In this work, three different datasets, such as the PIMA Indian dataset (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database), OhioT1DM Dataset (http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html), and Azure diabetic dataset(https://learn.microsoft.com/en-us/azure/open-datasets/dataset-catalog) information, are utilized. Each dataset has a set of information used to predict the patient's diabetic conditions using different attributes.

3.2.1.2 PIMA Indian Diabetic Dataset

The dataset comprises 768 patient records, including pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetic pedigree function, and age. The respective outcome is given to training the model along with these attributes. The trained model is used to predict the unknown data. Therefore, the dataset has 768 instances and nine attributes for analysing and classifying diabetic disease. From the collected information, 500 records are belonging to 0 classes and 268 records are belongs to 1 classes. The feature explanation is defined in Table 3.2.

Table 3.2: PIMA Indian Dataset-Feature Explanation

S.No	Features	Explanation
1	Glucose	The concentration of glucose in the plasma after two hours of an oral glucose tolerance test
2	Pregnancies	The total number of pregnancies
3	Skin Thickness	The thickness of the triceps skinfold (mm)
4	Blood Pressure	Blood pressure on the diastolic side (mm Hg)
5	Body Mass Index (BMI)	The body mass index can be calculated by taking a person's weight in kilograms and dividing by the height in meters squared.
6	Insulin	2-Hour serum insulin (mu U/ml)
7	Diabetes Pedigree Function	Pedigree Function
8	Age	In years
9	Outcome	Regardless of whether or not the patient has diabetes

According to the discussion, the sample PIMA Indian Dataset information is shown in Table 3.3.

Table 3.3: Sample PIMA Indian Dataset

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1

In addition to this, the distribution of the PIMA Indian Dataset is illustrated in Figure 3.2.

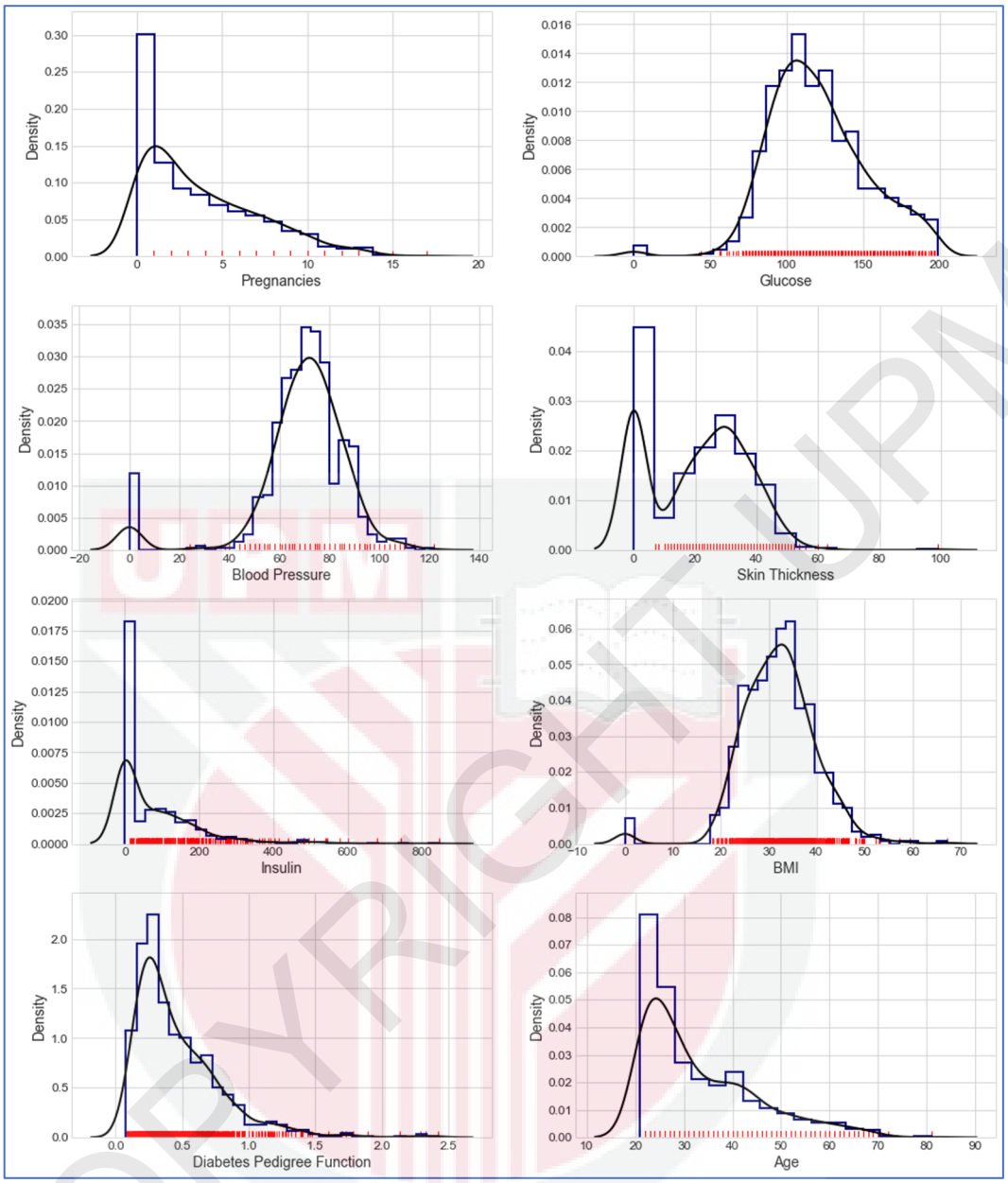


Figure 3.2: Distribution of PIMA Indian Diabetic Dataset

OhioT1DM Dataset

Ohio State University researchers dedicated to enhancing the lives of patients with type 1 diabetes can access the OhioT1DM dataset at http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html. Twelve participants with type 1 diabetes had the data covering eight weeks included. These patients were undergoing

CGM-guided insulin pump therapy. The Individuals gave information from physiological fitness bands, self-reports of life events, and blood glucose and insulin levels. Data from continuous glucose monitors (CGMs), self-monitoring of blood glucose (finger sticks), bolus and basal insulin doses, self-reported meal times with estimated carbohydrate intake, self-reported exercise, sleep, work, stress, and illness patterns, and physiological data from fitness bands are all included in the dataset. From the analysis, 721 instances belongs to class 0 and 88 samples are belongs to class 1. Table 3.4 details the characteristics and context of the OhioT1DM dataset.

Table 3.4: OhioTIDM Dataset Feature Explanation

S.No	Feature	Explanation
1	Patient	Patient ID and insulin type
2	Glucose Level	Frequent glucose monitoring every 5 minutes
3	Finger Stick	Self-assessment-based monitored blood glucose
4	Basal	The constant rate at which insulin for basal needs is injected into the body. The basal rate starts being applied at the timestamp (ts) that has been selected, and this will remain at that level until another basal rate is applied.
5	Temp Basal	Temporary basal rate
6	Bolus	Insulin is given to the patient, most frequently just before a meal or when the patient is experiencing high blood sugar levels.
7	Meal	Type of meal that the patient reports
8	Sleep	Assessing the sleep quality: poor-1, Fair-2, and Good-3
9	Work	A patient's working time is measured on a scale of 1 to 10.
10	Stressors	Self-reported stress time
11	Hypo Event	Time of the hypoglycaemic episode that was self-reported
12	Illness	Time of illness (self-reported)
13	Exercise	Time-duration of the patient exercise (rated from 1 to 10)
14	Basis Heart rate	Monitoring heart rate every 5 minutes
15	Basis GSR	The Galvanic Skin Response rate is monitored every 5 minutes
16	Basis Skin Temperature	Temperature is observed every 5 minutes.
17	Basis Steps	Counting patient movement (steps) for every 5 minutes
18	Basis Sleep	Sensor band reporting time of patient sleep
19	Acceleration	The acceleration magnitude value is computed for every minute

In addition, the distribution of the OhioTIDM dataset information is illustrated in Figure 3.3.

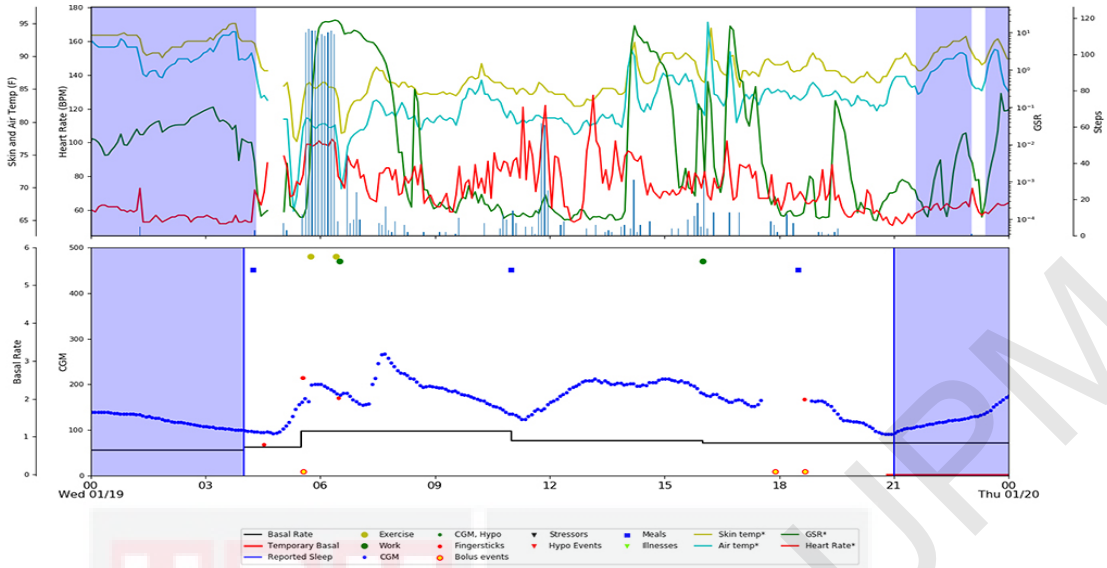


Figure 3.3: Distribution of OhioTIDM Dataset Information

Azure diabetic dataset

Ideal for learning how to implement machine learning methods, the Azure Diabetic dataset (https://learn.microsoft.com/en-us/azure/open-datasets/dataset-diabetes?tabs= azureml- opendatasets) contains 442 samples and ten features in which 268 samples are belongs to class 0 and 174 samples are depends on class 1. This data collection has 10 factors to consider, including age, BMI, sex, average blood pressure, and six different serum measures. Azure data set corresponds to the format of structured data or unstructured data; Common descriptive elements are metadata for features (name, data type (numeric, categorical), ID, etc.) It ranges from application specific variables with application specific information — demographics in healthcare or retail and timestamps or event specific metrics in IoT analytics. In order to work with these datasets, they are generally saved and accessed through various Azure services (such as Blob Storage or Data Lake), which allow for integration for preprocessing or modeling. Table 3.5 shows the sample Azure diabetes dataset.

Table 3.5: Sample Azure Diabetic Dataset Information

AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
59	2	32.1	101	157	93.2	38	4	4.8598	87	151
48	1	21.6	87	183	103.2	70	3	3.8918	69	75
72	2	30.5	93	156	93.6	41	4	4.6728	85	141
24	1	25.3	84	198	131.4	40	5	4.8903	89	206
50	1	23	101	192	125.4	52	4	4.2905	80	135
23	1	22.6	89	139	64.8	61	2	4.1897	68	97
36	2	22	90	160	99.6	50	3	3.9512	82	138
66	2	26.2	114	255	185	56	4.55	4.2485	92	63
60	2	32.1	83	179	119.4	42	4	4.4773	94	110
29	1	30	85	180	93.4	43	4	5.3845	88	310

Information on people with diabetes is analysed using data mining and improved machine learning methods. The optimized techniques are utilized to resolve the data assimilation, availability, and classification issues. The diabetic class in the above datasets such as Pima Indian and Azure dataset is imbalanced, meaning there will be more samples of one class than the other, so the machine learning algorithm may always predict the most common class. The dataset is typically imbalanced towards a larger ratio of healthy (non-diabetic) to diabetic samples. To confront this situation without applying some techniques like resampling or cost-sensitive learning, it has no support that gives up the ability to correctly predict the minority cases (diabetic ones) when facing class imbalance. For example, in a normal class distribution, the healthy class (non-diabetic) would have 70-80% of the data points and diabetic cases would have 20-30%. Such imbalance may be a source of bias, resulting in decreased predictive performance of the minority class. The collected data consists of inconsistent information affecting the overall diabetic prediction efficiency. Therefore, the gathered health details must be explored with the help of the pre-processing technique. Then only the remaining algorithm procedures produce effective results. Then the overall working process of the flock-optimized-related diabetic detection process is illustrated in Figure 3.4.

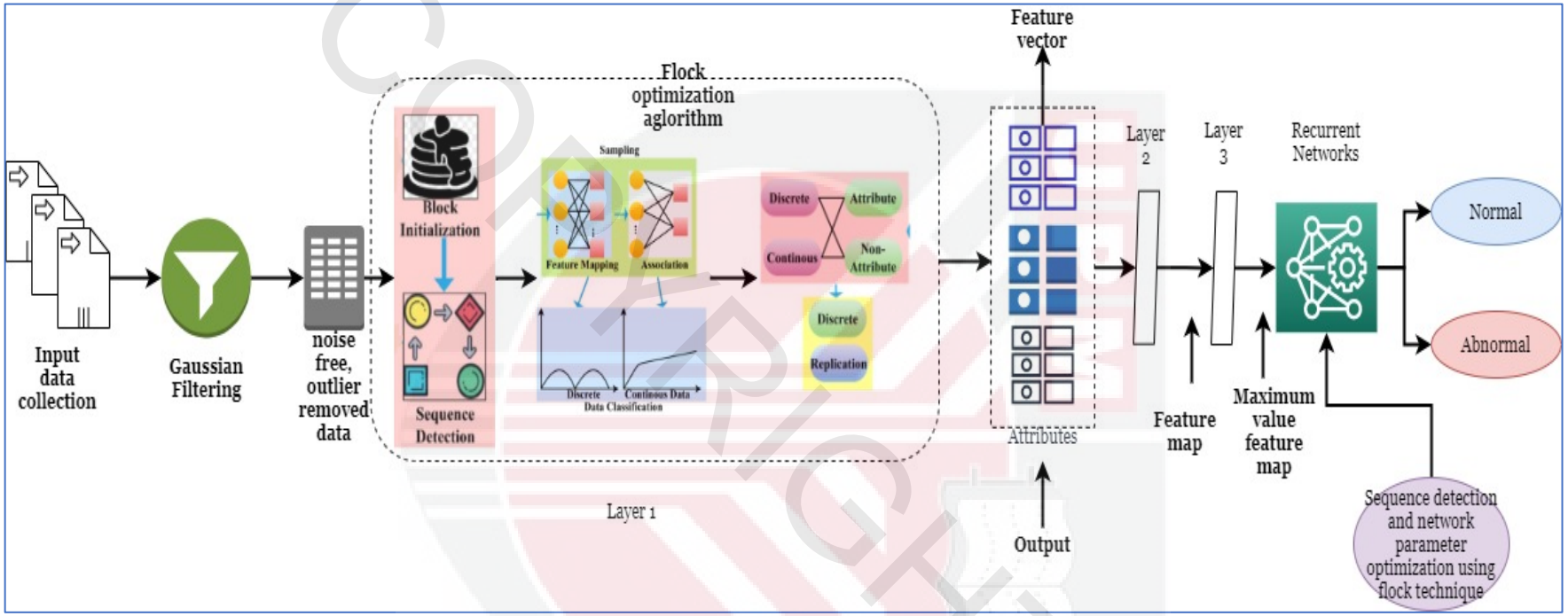


Figure 3.4: Flock Optimized related Diabetic detection Process

3.2.1.3 Preprocessing

This work uses the Gaussian Filter and normalization process to remove the noise from the collected diabetic dataset information. Then the overall working process illustrated in Figure 3.5.

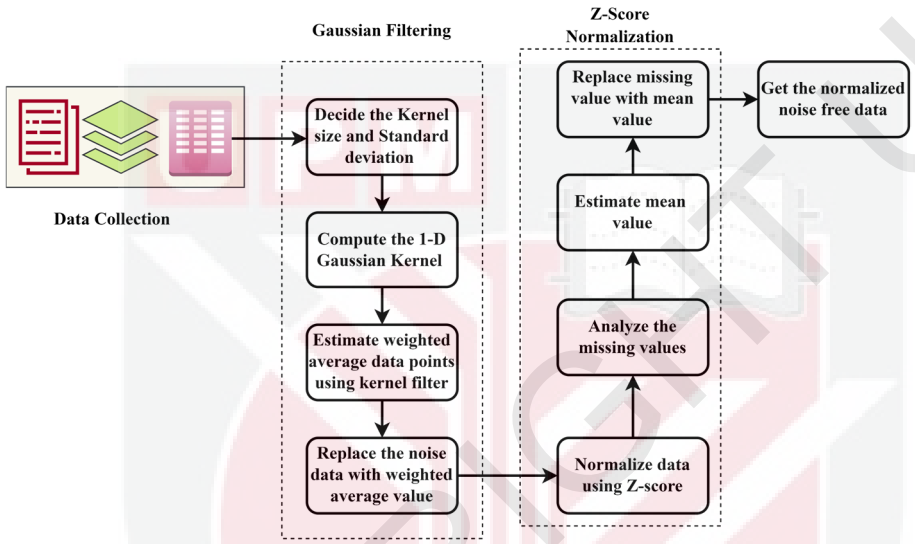


Figure 3.5: Working process of Gaussian Filtering and Normalization based data pre-processing

The Gaussian filter (Bishop et al. 2023) is one of the linear filters used to remove irrelevant, noisy, and inconsistent information. The high-frequency information has been filtered out as noise using a Gaussian distribution. The Gaussian filter is a type of low-pass filter that effectively gets rid of smeared information. The Gaussian filter is developed on the Odd-Sized Symmetric Kernel applied to data. This process determines the noise in every pixel, improving the overall classification efficiency. The Gaussian filter utilizes the kernel value or matrix to eliminate the noise from the data. The kernel values are mostly odd in numbers used to compute the noise information from the dataset. The Gaussian function (Nandan, D., et al. 2024) is estimated using equation (3.1).

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \quad (3.1)$$

Considering the dataset has a set of inputs that are defined as $I = (i_1, i_2, i_3, \ldots . i_n)$ that are D in dimension. The input I is convolved with the Gaussian function $G_\sigma(x,y)$ to get the output as filtered input that is defined as (fI). In equation (3.1), X and Y are defined as the coordinated value of inputs, and the standard deviation is defined as σ. The computed Gaussian filter value is utilized to replace the irrelevant and noisy information. The Gaussian filter successfully analyses the signal-based diabetic information. In this study Pima Indian Diabetic dataset is utilized to predict diabetic disease. Preprocessing of the Pima Indian Diabetes dataset is a crucial initial phase in the data analysis and machine learning pipeline. This dataset encompasses health-related attributes of Pima Indian women, coupled with the diabetic status. The pre-processing endeavours serve multiple pivotal objectives. Firstly, data cleaning procedures are applied to rectify missing values, address outliers, and rectify inconsistencies, ensuring the dataset's integrity. Secondly, feature selection techniques are employed to identify and retain the most relevant attributes, promoting efficient model performance and reducing computational complexity. Thirdly, feature scaling is executed to standardize attribute magnitudes, mitigating issues stemming from disparate scales and facilitating faster algorithm convergence. Categorical variables are encoded numerically, rendering them compatible with machine learning algorithms. Normalization is conducted to establish a common scale, typically between 0 and 1, enhancing the suitability of certain algorithms for data analysis. Strategies to rectify class imbalance, such as oversampling or under-sampling, are implemented if necessary. Redundant or highly correlated features are pruned, streamlining the model and potentially enhancing interpretability. Temporal data, if

present, may necessitate specialized pre-processing techniques like lagging or windowing. Ultimately, the primary goal of pre-processing the Pima Indian Diabetes dataset is to transform the raw data into a well-structured, clean, and appropriately formatted dataset. This refined dataset is a foundation for subsequent analysis, modelling, and predictive tasks related to diabetes detection or other pertinent objectives. The specific pre-processing steps depend on the analysis's ultimate aims and the machine learning algorithms being employed.

Outlier Detection

The first step is to analyse the dataset to remove the outlier information. Z-score normalization (Geem, D., et al. 2024) is a data pre-processing technique that transforms numerical data into a standard scale, making comparing and analysing variables with different units and ranges easier. This technique is particularly useful in identifying and predicting outliers within a dataset. Z-score normalization is advantageous for outlier detection because of considering the distribution of the data and normalizes to a common scale. This helps to identify outliers relative to the variability of the dataset, making the technique more robust compared to fixed threshold-based methods.

Additionally, z-score normalization (Kholwal, R. (2023)) maintains the shape of the original distribution, which can be useful for visualizing and interpreting the data. The Z-score value is initially computed to transform the data into the standard scale. This determines the z-score for each observation in a dataset, which measures the number of standard deviations from the mean. This technique is particularly helpful for outlier detection. The z-score value is computed using equation (3.2).

$$Z = \frac{X-\mu}{\sigma} \tag{3.2}$$

In equation (3.2), the data point z-score value is represented as Z; particular data is X, the mean value of the dataset is represented as μ, and the dataset standard deviation value is σ. The data is transformed by removing the mean and dividing by the standard deviation after the z-score has been computed for each data point in a given feature. This normalizes the data to a zero-centred scale based on the standard deviation. Removing the mean and dividing by the standard deviation are two common methods for normalizing data. The modified data have a mean of 0 and a standard deviation of 1 after normalization. This facilitates the examination and comparison of variables on varying scales. The results of z-score normalization have a mean of 0 and a standard deviation of 1 by definition. Data points far from the mean (i.e., those with high absolute z-scores) are considered potential outliers. Outliers are typically defined as data points with z-scores above a certain threshold, often set at ±2.5 standard deviations from the mean. A common threshold for outliers is ±2.5 SD, which strikes a balance between sensitivity and robustness. It grants few exclusions, thus will be very conservative on usage of valid data, while providing a good separation of the extreme points of normally distributed data sets. The Berger, A., & Kiefer, M. (2021) studies emphasize that thresholds of the order of 2 to 3 standard deviations are usable across different contexts and ensure low Type-I error rates without destroying the statistics of the dataset to be analyzed. Long analyses of reaction time and other datasets show that such thresholds help maintain data integrity without throwing away data or compensating too much for outliers. The normalization process is performed according to equation (3.3).

$$Normalization\ (x, cut) = \begin{cases} x_i & if\ |x_i - \mu| < \sigma * cut \\ mode\ (x) & else \end{cases} \quad (3.3)$$

In equation (3.3), x_i is defined as the input in the dataset, the mean value of the dataset is defined as μ, and the standard deviation of the input value is denoted as σ. Then missing value is computed, and the missing values are replaced with the help of the mean value. The pre-processing procedure improves the overall data representation and maximizes the prediction accuracy. Here, the sample analysis of the Pima Indian dataset pre-processing procedure is explained. For demonstration purposes, let's consider a simplified version of the dataset with just two features: "Glucose Level" and "BMI" (Body Mass Index). Considering Table 3.6 data points for these two features:

Table 3.6: Sample Data points for outlier Analysis

Glucose Level	BMI
120	32
140	26
160	28
100	25
180	35

Step 1: Estimate the Mean and Standard Deviation

Initially, compute the mean (μ) and standard deviation (σ) value for every feature (Glucose level and BMI) in the dataset. The mean value of glucose $(\mu_{glucose})$ is computed as follows.

$$\mu_{glucose} = \frac{(120+140+160+100+180)}{5} = 140 \quad (3.4)$$

From the computed mean value, the standard deviation is estimated $\sigma_{glucose} = 28.28$, which is computed as follows.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{3.5}$$

$$\sigma = \sqrt{\frac{(120-140)^2 + \ldots\ldots + (180-140)^2}{5}}$$

$$\sigma = \sqrt{\frac{4000}{5}} = \sqrt{800} = 28.28$$

Step 2: Compute the z-score and normalized values

From the computed mean and standard deviation values, the z-score and normalized values are computed as follows.

$$z - score\left(Z_{glucose}\right) = \frac{(Glucose\ level - \mu_{glucose})}{\sigma_{glucose}} \tag{3.6}$$

$$Normalized\ (Glucose_{normalized}) = \frac{(Glucose\ level - \mu_{glucose})}{\sigma_{glucose}} \tag{3.7}$$

Similarly, BMI and other features are normalized to get the outlier information. During the analysis, 2.5 as utilized as the threshold value to predict and control the outliers. According to the analysis of various researches, 2.5 is selected as threshold value due to lies within the standard deviation from the mean (Schwarzhans, F., et al.2024). Then the computed results are shown in Table 3.7.

Table 3.7: Numerical Analysis of the Normalization

Glucose level	Z-score (glucose)	Absolute z-sore (glucose)>2.5	BMI	Z-score (BMI)	Absolute z-score (BMI)>2.5
120	-0.8944	No	32	0.6562	No
140	0	No	26	-1.3125	No
160	0.8944	No	28	-0.5208	No
100	-2.2361	No	25	-1.7812	No
180	1.2361	No	35	2.958	Yes

Based on the comparison, the data point with a BMI of 35 has an absolute z-score for BMI greater than 2.5, indicating that as a potential outlier according to the chosen threshold. If the dataset has any missing values, the mean values are computed to replace the particular features from the dataset. This process is repeated for all features in the dataset to improve the overall data analysis and classification process. According to the noise removal and normalization process, the distribution of the dataset is denoted in Figure 3.6.

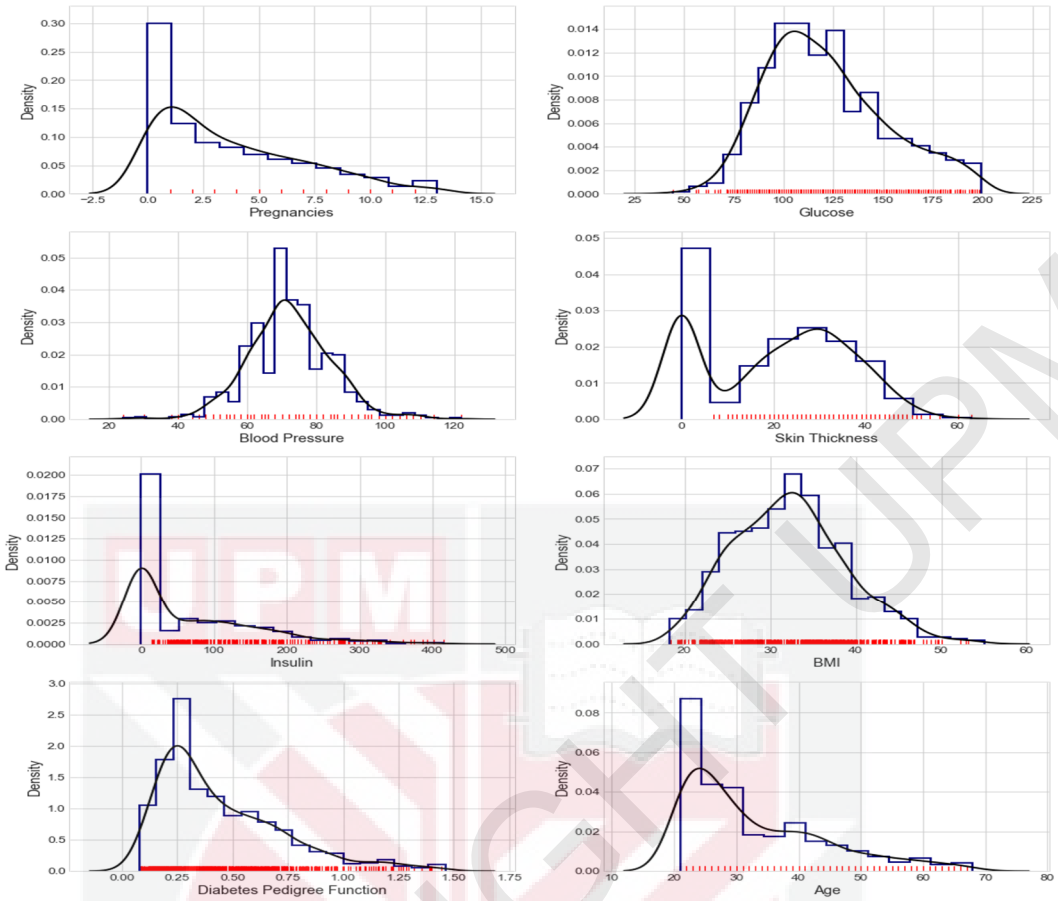


Figure 3.6: Normalization Process based Data Distribution Representation

In addition, the comparison of raw, filtered and normalized data of three datasets are shown in Table 3.8.

Table 3.8: Description of Dataset (raw, filtering and normalization)

Dataset	Raw data	Attribute range	Filtered Data	Attribute Range	Normalized data	Attribute Range
PIMA	768	[0,200]	768	[0.1,150]	768	[-1.5,2.5]
Ohiho	1010	[0,300]	1000	[0.2,250]	1000	[-2.0,3.0]
Azure	500	[0,250]	490	[0.15,200]	490	[-1.8,2.0]

3.2.1.4 Deep Learning for feature and data association classification

The next critical stage involves feature extraction using the assimilated deep learning model for flock optimization. Data availability complicates the task of categorizing a vast amount of data. In addition, data assimilation requires identifying the optimal solutions from the set of solutions. The model uses different parameters and interpolates sparse observation to classify the input details during the classification process. This work uses three information sets to assimilate and improve the diabetic classification process. The assimilated data may be large in dimension, creating the computation complexity and causes to the error rate. The error rate and other research classification issues are overcome by applying the flock optimization deep learning model.

The research uses the three statistical information-related datasets; therefore, the system faces the Bayesian Estimation Problem (BSF). The term Bayesian filtering refers to the process of recovering the hidden states of a dynamic system by combining the prediction information with the observed data according to the principles of Bayes' Theorem. This revises the estimates of the state as new observations are accessed, iteratively enhancing state estimates according to the probability distribution based on the observations made. BSF is widely applied to the systems where the properties of the state are obscured by uncertainty or noise in diabetic data classification. Bayesian estimation is a robust statistical framework for estimating model parameters by combining previous knowledge with evidence from the same or similar models. Bayes' theorem, a cornerstone of probability theory, is central to this method. This theorem provides a mathematical connection between initial assumptions about the parameters,

the probability of observing the data given those parameters, and the resultant revised assumptions reflected in the posterior distribution. Prior distributions, which capture early assumptions about the parameters, are established as a first step in the Bayesian estimation process. The likelihood function is then developed to characterize the probability of the observed data given the model. Bayes' theorem harmonizes these parts to produce the posterior distribution. This revised distribution summarizes researcher improved knowledge of the parameters after accounting for new information. The ability of Bayesian estimation to handle uncertainty consistently stands out since this can provide probabilistic distributions encompassing the space of conceivable parameter values. This makes useful in cases where either data is few or models are intricate. Inferences are drawn from the posterior distribution, typically producing point estimates or intervals that sum up the uncertainty in the parameters. The choice of priors, however, can have a substantial impact on the outcomes, making picking them difficult. Complex models can also provide difficulties due to the high processing requirements.

Generally, the data analysis process uses the Recursive Bayesian Estimation to compute the probability value of every input. According to the probability value, each data is frequently examined to improve the classification efficiency. The probability computation uses the mean and standard deviation value, which is more relevant to the Gaussian Filter. Therefore, the noise data is removed during the data analysis process. The preprocessed information is arranged in the search space S to analyse the features according to the different characteristics, such as sex, insulin, glucose level, BMI, age, etc., for every data; features are computed that examine the relationship between each data and diabetic disease. The flock optimization algorithm is applied during this process to examine the relationship between the data in S.

3.2.1.5 Sampling-based Flock Optimization and Deep Learning Assimilated for feature mapping

Flock optimization with sample-based learning (deep learning) to facilitate feature mapping in a data-driven environment. This novel method improves complex data comprehension and representation by combining optimization inspired by swarm intelligence with feature extraction using neural networks. The concept of "flock optimization" is based on the efficient group dynamics seen in flocks of birds and other animals. In order to produce optimization algorithms, this models the interactions and cooperation seen in these communities. A flock of potential solutions (individuals) is generated as part of the optimization process. These individuals explore the space of possible solutions through interactions, converge on the best ones. Flock optimization is used in feature mapping to find the optimum features to map a high-dimensional dataset onto its underlying structures and relationships. The foundation of this architecture lies in the fusion of sampling-based flock optimization and deep learning techniques. The former draws inspiration from the cooperative behaviours observed in natural flocks of birds, translating these principles into an optimization strategy. This strategy involves generating a population of candidate solutions that interact to navigate towards optimal outcomes, here directed towards identifying salient features in high-dimensional data. Concurrently, deep learning, exemplified by neural networks, automatically extracts intricate features from raw or preprocessed data. The crux of this integration emerges from the mutual interaction between these two components. The insights gleaned from sampling-based flock optimization guide the training and refinement of the deep learning model, while the learned features from deep learning expedite and inform the optimization process. This iterative feedback loop leads to an enriched understanding of data through enhanced feature representation.

Flock Optimization is a meta-heuristic approach that works according to the starling bird's activity. The starling bird is similar to the other animal's interesting group dynamics. The bird movements have rules used to make the relationship between each other and help generate collective behaviour. Each flock's starling motion is observed to explore the neighbouring bird's motion and response. The birds respond to other birds according to the distance measure (topological distance). During this process, the flock optimization model uses every bird in the search space, and behaviour is observed to predict the best solution.

Flock optimization denoted as FlockOpt, consists of a flock of flock birds. The behaviour of the birds is monitored with the help of several cameras that are used to reconstruct the bird's behaviour and starling. This research uses the seven neighbouring birds monitored for every bird according to the motion during the feature analysis. This method is based on flocking behaviour, which allows each individual to update its motion relative to the nearest individuals. While monitoring seven neighbouring birds, the newly introduced method captures interactions at local level which help in better feature extraction of the collective motion. The recordings made with cameras allow for accurate behavioural replication for subsequent analyses. The bird distributions are varied, requiring the distance computation to observe every bird activity in search space. The algorithm has three basic steps while identifying the optimal solution. As said, the search space S has several flocks of birds denoted as i,j,….n. Considering the agent i and j then, the distance is computed d_{ij} while selecting the optimal solution. Three basic steps are listed as follows.

- Select the individual i from search space S randomly
- Choose the updated partner j for individual i with $P_j \sim {}^{1}/_{d_{ij}}$
- Update the agent i velocity and position according to the computed distance value d_{ij} (the distance value may be varied as large, medium, and small).

According to the above steps, agent velocity is updated based on the distance measure. If the distance between the agents is small, the birds move each other to eliminate the collision. Suppose the birds having the medium or intermediate distance value i allocates velocity to the i to j. The distance between i and j has far; then, each bird is attracted to the other, and flock cohesion is maintained. The computed three distances have three parameters such as radius repulsion (rR), radius attraction (rA), and radius orientation (rO). According to the bode model, d_{ij} is maximum compared to (rA), then each agent does not have any interaction. The Bode model (Izci, D., et al., 2023, N. W. F. Bode et al. 2011) requires many updates on each individual to produce accurate simulation outputs. Before each simulation output, each individual's instantaneous velocity updates (repulsion, alignment, and attraction) are averaged together to create a more accurate picture. The computer simulations of enormous populations can recreate statistical results comparable to those obtained from the starling flocks. Because of this, data produces a flock that moves inside a framework based on topological distance rather than metric distance. This allows for more accurate tracking of the flock's movements.

The Bode Model inspiration was utilized in the flock optimization process to select the best solution in the search space. The flock birds are moved in the search space depending on the measurement. As discussed above, three radius measures are utilized

to select the optimal solution. Three radius parameters are utilized in the selected agents to repulse the partner, align with the partner and attract each other (depending on the distance). In addition to the Bode model, the pbest term is utilized in the model to improve the solution detection efficiency. The pbest term resolves the optimization problem while selecting the agents in the search space. The utilized pbest term works concept of the Particle Swarm Optimization (PSO) that update the agent velocity. The updating process defined in equation (3.8).

$$pbest = xp_i - x_i \tag{3.8}$$

In equation (3.8), the value x_i is used to reflect the agent's current position, while the value xp_i is used to indicating the ith optimum measurement site. The pbest solution that was found is placed to use in trying to determine the search direction within the search space. This has been decided to update the positions and velocities of the agents during each phase of the search process to make the sequence detection procedure less complex. As was previously mentioned, the agent i velocity, which is indicated by the notation v_i, needs to be updated for each search, and lastly, the average value of the velocity increments needs to be computed. The value of the agent's position is then updated following the computation of the velocity value. During the process of looking, agents are selected at random and replaced with other agents as needed. The criterion $d_{ij} < rA$ needs to be met to choose the partner j for every agent i. In this case, the repulsion is represented by the symbol rA, while the distance between the agents is marked by the symbol d_{ij}. In this technique, the selection of the neighboring agent j is accomplished by the use of the roulette wheel selection process. The probability value can be found by applying the equation (3.9).

$$p_j = \frac{(1/d_{ij})}{\sum_k (1/d_{ij})} \tag{3.9}$$

Following equation 3.9, the neighbouring agents are calculated while staying inside the bounds of the rA for agent i. This procedure iterated over every piece of data contained inside the search space. The amount of time spent looking needs to be more than the total number of people searched in the area. In most cases, population time is utilized in searching for characteristics inside the search space. During the search procedure, the pbest solution is determined with the help of the fitness function. Following the proposed solution, a succession of features can be identified in the data list. Analyzing the relationship between the data is necessary to discover the important characteristics. Sampling strategies are implemented throughout the investigation to facilitate an overall improvement in the relationship detection process. The clusters are developed based on the similarities in the data after the sampling procedure has investigated all of the dataset's characteristics. The primary objective of cluster sampling is to enhance disease prediction accuracy, simplify implementation, provide convenient access, and reduce computing time. In addition, the sampling method reduces the variability and ensures the project's feasibility is achieved. The cluster samples are selected first in cluster sampling, and the elements that will comprise the clusters are decided upon.

3.2.1.6 Sampling

Cluster sampling is a type of probability sampling in which the population is divided into various groups or clusters to conduct research. Next, the researchers will use either a basic random or an ordered random selection procedure to choose groups at random

for data collection and analysis. Cluster sampling is applied to mutually homogeneous activities to analyse the statistical information. Cluster sampling consumes minimum time, cost, convenient access, ease of implementation, and data accuracy. These advantages are the main reason the sampling technique was incorporated to improve the overall diabetic data analysis.

This work uses Simple Random Sampling (SRS) to analyse the diabetic data to predict the patient's health condition. Researchers use simple random sampling (SRS), a form of probability sampling, to randomly select participants from a community. Everyone in the population has an equal chance of being selected out of those who are there. This strategy typically results in samples representative of the whole and unbiased. A sampling method known as simple random sampling helps to ensure that the sample is representative of the population as a whole. The procedure takes substantially more samples from major subpopulations than from lesser subpopulations on a more consistent basis. The simple random sampling method has a few basic steps that are listed as follows.

- Describe the population to analyse the data for selecting the optimal data.
- Generate the list of members to the respective population.
- Allocate the random number to every member.
- Select participants using a random number generator until achieving the desired sample size.

According to the simple procedure, the samples are analysed in the search space, and the relationship between the data is examined. The sampling-based SRS is the most straightforward approach to taking an objective sample from a population. The researchers do not require any further information regarding the population, its

subpopulations, or the characteristics of the population; however, do require a list of the entire population.

On the other hand, more complicated sampling techniques require researchers to understand the population's features. After that, to prepare for the sampling process, the population is first segmented into clusters using the information previously gathered. Then several operations are carried out using SRS; select subjects randomly from the list until its sufficient numbers.

Analysts who need to utilize a sample to deduce the characteristics of a population will find that simple random sampling is an excellent method because this tends to create unbiased samples that are reflective of the population (i.e., inferential statistics). A study's internal and external validity of a study can both be improved by having a sample representative of the whole. After a simple random sample, statistical hypothesis testing is utilized to derive inferences about the population based on the sample.

Let's consider that the dataset contains N different populations and that throughout the sampling procedure, n different clusters were chosen from each of those populations. Following the selection of the cluster, a process known as Simple Random Sampling (SRS) is utilized to determine the cluster's components. The members are chosen for the SRS process from the collected samples carried out randomly. Because of this factor, each subset will have the same probability value during the selection process. This is referred to as unbiased sampling. Equal Probability Sampling is another name for the SRS procedure, which ensures that each population member has the same

chance of being selected (EPSEM). The data list is analysed in the first step of the sampling procedure, after which the samples with the same probability value are extracted from the list and grouped. This procedure for sampling was carried out multiple times for each k-sized entry in the data list. The computing cost and the amount of time spent searching for data can be reduced by selecting a multi-stage sampling method with a recurrent selection of N number of clusters and n number of members. The relationship between the two data sets can be efficiently predicted using the computed probability value and uniform distributions. After that, each sample is analysed, and sequence attribute detection is performed following the flock optimization procedure. Because the interrelationship data only identifies the most relevant patterns associated with attributes, detecting the attributes and relationships helps resolve the computation's complexity. When determining the feature vector, the clustered inputs are processed through convolutions that include multiple layers (F). The Flock-based CNN-optimised feature modelling approach is introduced for creating the Diabetic Data Classification (DDC). PIMA Indian dataset information is utilized during the analysis to identify diabetic disease. Diabetes can be classified into diabetes due to β cells, diabetes due to progressive insulin, gestational diabetes mellitus and diabetes due to other causes. By these types, the classification process is done by the DDC. The patients are classified according to the types and health data. The DDC is used to classify the patients depending on the diabetes types and health issues which are represented in the diabetic and health data.

This Convolution Neural Network (CNN) is used to increase the dataset of the patients and to improve the process of classification accordingly. Neural Networks rely on the load of the data provided by the diabetic and health information sent by the analysing

data. As this has the multilayer perceptron with correct information on the diabetic data of the patients produced by the health data. This is also used to reduce the overfitting data during the process of finding the disease accuracy. This classification done by the DDC with the CNN helps in the reduction of the complexity using smaller and simpler patterns to detect the accuracy. This is used to classify diabetes according to its types and as per the patient as its done by the output of the diabetic data which is produced by the analysing data. The process of classification done by the DDC with CNN is explained by the following equations given below. The convolution procedure is carried out following the equation (3.10).

$$L_l(x) \leftarrow L_l(x) - \forall \frac{\delta D}{\delta L_l(x)} T_l(x) \tag{3.10}$$

Where $T_l(x)$ is denoted as the estimation of the classification process done by the DDC, $\forall \frac{\delta D}{\delta L_l(x)}$ is denoted as the data collected from the output of the diabetic data, which is used in the classification process. Now the sequence is determined from the classification process which is done by DDC. The sequence is an enumerated collection of the patient's health in which repetitions are allowed in the classification process. This contains the patient's health and the disease data used in the classification process. The amount of the process done by the classification is defining the sequence of the procedure. By this output of the sequence, the optimization problems can be reduced by classifying the diabetic data by using DDC with the help of CNN. Figure 3.7 presents the data extraction post the classification for sequence detection.

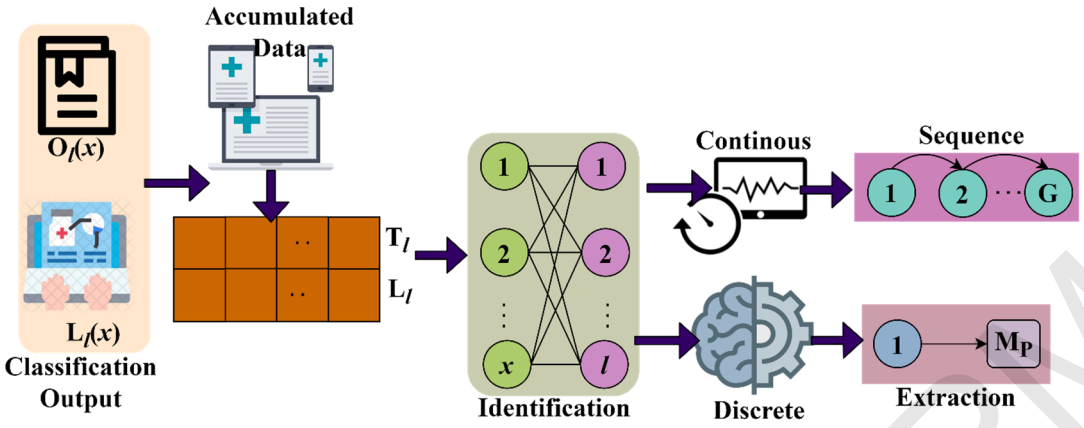


Figure 3.7: Process of Sequence Detection

The classification outputs $O_l(x)$ and $L_l(x)$ are used for validating $\forall \frac{\delta_D}{\delta L_l(x)}$. This validation is performed for $(x \times l)$ combinations such that continuous and discrete instances are analyzed. Considering the T_l to L_l matching, the continuous 1 to G and discrete M_p is identified. This is required for extracting $M_p \in (1, G)$ for new sequences (Figure. 3.7). The convergence difficulty can be sidestepped by using the classification process to perform the sequence in order. This DDC system employs CNN to aid with diabetic data classification, and this can sequentially order classification data to avert issues. When this comes to the categorization process based on employing neural networks, the order of the diabetic diseases counts. Multiple instances of the same diabetic order can occur in any given sequence. The classification process's function can then be used as the sequence's starting point. Problems, both current and potential, can be identified and resolved with the use of the sequence's identifying string. The DDC employs the CNN for the classification process, which is to establishes the sequence. This is a tally of diabetes cases and patient health records organized by diabetes diagnosis. Following are some equations that help explain how the DDC uses CNN to classify data and then provide that data in sequence.

$$L_l(x) = \left(\frac{L_l(x)}{G}\right) G \qquad (G = 1^n) \tag{3.11}$$

In equation (3.11) G represents the predicted sequence based on DDC classification using CNN. After the sequence has played out, the convergence problem must now be identified. The flock-based DDC and CNN-assisted condition determination. This sequence execution-independent convergence problem does not initiate execution termination. Additionally, this offers incrementally varied outcomes at each stage of processing. DDC combined with CNN can significantly reduce and sequence discontinuities in the diabetes data output by the classification method may cause a convergence difficulty. Using the data sequence in which the convergence is detected, the existing diabetes data can be categorized according to diabetes severity using the flock optimization technique (J. Hereford and C. Blum, 2011). The algorithm is applied to the problem of optimizing the series of diabetes illness solutions by a specified measure. DDC combined with CNN's categorization abilities can also be used to jumpstart the solution improvement process. Better solutions are updated using the technique to prevent issues with convergence. The following equations demonstrate how to identify the sequence process problem using the classification approach and to solve.

$$M_l(x) = \begin{cases} 1 - G & M_l(x) \geq 1 \\ -1 & M_l(x) < -1 \\ \left(\frac{M_l(x)}{G}\right) G & M_l(x) \epsilon (-1,1) \end{cases} \tag{3.12}$$

$$M_P = \frac{\sum_{NP}(1 - R_{lP})}{R_{lP}} \tag{3.13}$$

In equation (3.12) $M_I(x)$ is denoted as the determination of the problems, M_P is denoted as the data extracted from the sequence, $\frac{\sum_{NP}(1-R_{IP})}{R_{IP}}$ is denoted as the estimation of the problem-reduction process. The health data is currently being filtered using a Gaussian distribution. Health data is filtered to remove unnecessary information to improve diabetes prediction. Over-fitting problems are mitigated by using a Gaussian filtering method on the obtained data to remove extraneous details. In order to distinguish between typical and aberrant characteristics, a recurrent neural method is used. Predictions of diabetes are made using health records and Gaussian filtering. Figure 3.8 depicts the filtering procedure.

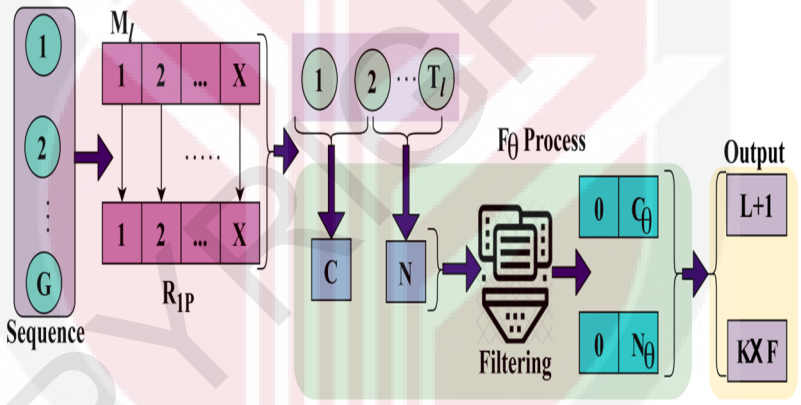


Figure 3.8: Illustration of Filtering Process

Gaussian filtering's disease prediction technique is applied to health data in order to make inferences about the likelihood of diabetes, and this also aids in minimizing problem fitting and also aids in the intricate forecasting of diseases with hidden prevailing characteristics. In order to avoid convergence and optimization issues, this filtering helps identify the forecasts and significance of the diseases. This also put to use to head off the development of diabetes-related health problems and difficult to make accurate predictions about the disease because of the danger that poses to those

who have yet to manage to cure or keep under control. The accuracy of diabetic disease detection is improved by using these Gaussian filtering predictions. The following equation describes how Gaussian filtering is used to make predictions from health records.

$$F_\theta = \sum[(L+1) \times K \times F] \times \left[\frac{C}{C_\theta}\right] \times \left[\frac{N}{N_\theta}\right] \quad (3.14)$$

$$F_\theta = \sum[(L+1) \times K \times F] \times \left[\frac{C \times B}{C_\theta}\right] \times \left[\frac{N}{N_\theta}\right] \quad (3.15)$$

Where F_θ is denoted as the estimation of the Gaussian filtering, $\left[\frac{C}{C_\theta}\right]$ is denoted as the predictions of diabetic disease, $\left[\frac{N}{N_\theta}\right]$ is denoted as the detection of optimization problems. Gaussian filtering's relevance analysis is currently underway. Its primary use is diagnostic and can detect the presence of diabetes in a patient. Filtering can be used to isolate patient health records that pertain to diabetes. With this method, unnecessary patient health records can be discarded. Overfitting, convergent issues, and irrelevant data can all be filtered out with this method. In addition to improving the accuracy of diagnosing diabetes, this method also removes the possibility of data forgery on the part of healthcare data. To further understand how Gaussian filtering works and how this might be used to eliminate noise in data, consider the equations below.

$$C_\theta = \frac{N - N_\theta}{N} \times (L \times K \times H) \quad (3.16)$$

$$F_w = \sum Q_l (1 - R_{lp}) \quad (3.17)$$

Where $\frac{N-N_\theta}{N}$ is denoted as the estimation of the relevant data, F_w is denoted as the calculation of irrelevant data, $\sum Q_l(1-R_{lp})$ is denoted as the elimination of irrelevant data.

The presence or absence of diabetes can now be deduced from the results of the Gaussian filtering procedure. In order to distinguish between typical and aberrant characteristics, a recurrent neural method is used. The following equation describes this procedure:

$$C_\lambda = \sum_{P_1}(W) \times L \times K \times \left[\frac{N}{N_\theta}\right] + \sum_{P_2}(W) \times \left[\frac{C \times B}{C_\theta}\right] \tag{3.18}$$

Where C_λ is denoted as the estimation of the diabetic status, (P_1, P_2) is denoted as the output of the filtering process. The CNN process illustration is portrayed in Figure 3.9

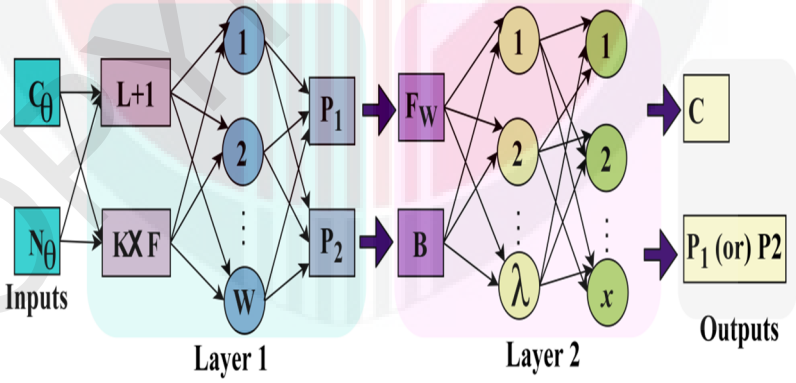


Figure 3.9: CNN process

According to the above process, the output is obtained with the help of the sequence of layers in the convolution networks. Then the output is derived using below equation (3.19).

$$z[m]=\sum_{i=-l}^{l} x[i].\delta[m-i] \quad (3.19)$$

In equation (3.19), $x[i]$ Input, kernel - δ $(size\ 8)$, convolution result is z with index m, in the first layer. Then the computed convolution z[m] is transferred into the next convolution layer with a pooling function to obtain the feature map (y^*) which is obtained after performing down-sampling in that convolution layer. The maximum feature map (Y) is obtained by examining (y^*) and the computed features (Y) are fed into the feature selection process to minimize the computation difficulties.

3.2.1.7 Deep Learning Flock Optimization for Prediction

The last step of this work is a diabetic prediction done with the help of the Long Short Term Memory Neural Network (LSTM). This is one of the neural networks but has feedback connections that are used to improve the overall data analysis. Then the overall diabetic classification process using optimized learning process is illustrated in Figure 3.10

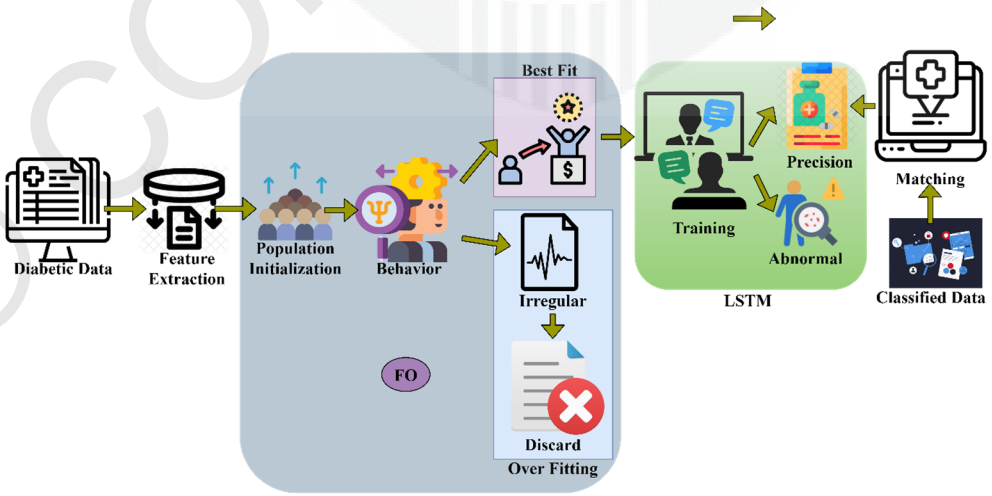


Figure 3.10: Structure of Deep Learning Flock Optimization for Prediction

Estimates and classifications of diabetes data are used to determine an optimal course of action. Features are derived from the diabetic data and used in the optimization procedure. The input for the initialization of the population consists of these characteristics. Its inhabitants are known as lemur monkeys. And from this, inferences about the data's behaviour can be made. The behaviour allows for the detection of best-fit and overfitting data which has been determined that the jump-up is the optimal strategy. The information should be presented in a logical order with no overlap. Overfitting behaviour in data features is considered abnormal and is eliminated. Data meanings will vary, and there will be some overlap, therefore the replicated data will be erratic. The primary purpose is to identify the most pertinent informational characteristic among the highest number of flock members displaying the same type of leap-up behaviour. High-precision and unusual features can be classified with the use of this output, which is fed into deep learning to train exemplars. Existing categorized data that best fits its attributes is used to calculate the accuracy. If the data was less predictable than expected, the operation must be redone with new goals, at which time the leap-up pattern is employed to remove the overfitting parameters. This method eliminates the potential for accuracy loss. Here, the Flock Optimization (FO) method is applied to a dataset including information about diabetes. The FO optimization procedure begins with feature extraction from the diabetic data. In order to identify accurate, high-precision data without duplications or mistakes, characteristics are extracted for use in classification. Information required for classification is elicited by diabetes disease, and the output of the data features will be used as input data for the next stage of population commencing. Using the equation below, need to understand how features are extracted from the diabetes data that will be used as input.

$$\varphi_{\sigma} = \frac{1}{M}\sum_{i=1}^{M} G\left(b^{i}, f(a^{i}, \sigma)\right) \tag{3.20}$$

Where φ_{σ} is represented as the data from the diabetes disease, a^{i}, b^{i} are indicated as the features of the data from the diabetes disease, G is signified as the diabetes data, M is represented as the process of extraction, f is symbolized as the input data. The extracted features are fed into the Long Short Term Memory network to identify the normal and abnormal data. The Long Short Term Memory handles the sequence of data, which is specific characteristics that maximize the data prediction. The network has a memory unit to store every processing input. The effective utilization of memory units, different gates, and activation functions causes the Long Short Term Memory to be applied in different applications. Healthcare systems, video games, robot control, speech recognition, handwriting recognition, and machine translations are important applications of the Long Short Term Memory approach.

The Long Short Term Memory network has a set of nodes, connections, and parameters. For every data training, parameters (weight and bias) are frequently updated, minimizing the overall classification deviations and improving the prediction accuracy. During training, the network frequently updates these parameters, and the updated inputs are stored in the memory unit. The Long Short Term Memory has long-term memory units for storing every timestamp information. The Long Short Term Memory has cell, input, forget, and output gates. The cells save the entire processing inputs at timestamps and regulate the network flow. The input gate receives the information from the feature extraction step. The received input is processed by forget gate that determines the importance of the feature and also forget to compare the inputs with the previous state output value to determine the status of the information. If the gate returns zero as the output, then the information is discarded from the list, and one

is saved in the memory for further analysis. The stored input is fed into the output gate that uses the activation function to compute the output value (0 or 1). The output value represents the diabetic disease status (abnormal or normal). The Long Short Term Memory can process the sequence of data, which means the network can effectively identify the present and future inputs. In addition, the network is designed to address the vanishing gradient problem. The gradient problem is identified during the feature training. This gradient issue is addressed by updating the network parameters, such as weight and bias. The frequent updating of network parameters minimizes the computation difficulties and misclassification error rate. According to the discussions, the Long Short Term Memory network structure is illustrated in Figure 3.11

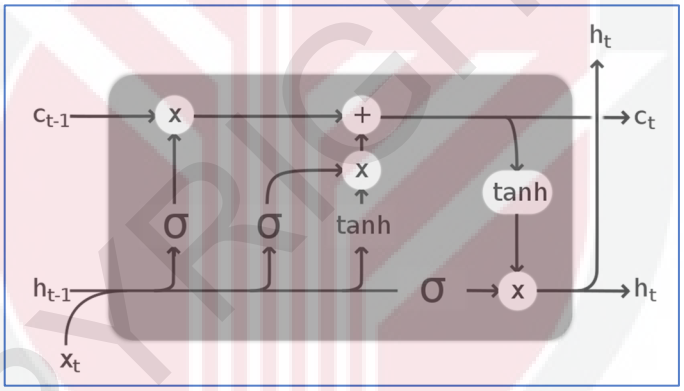


Figure 3.11: Long Short Term Memory Network Architecture

Figure 3.11, therefore, presents Long Short Term Memory, or long short-term memory. Its design has nearly solved the vanishing gradient problem, but the training model has remained the same. In addition to handling noise, dispersed representations, and continuous data, Long Short Term Memory's are useful for sometimes closing lengthy time gaps. Unlike the hidden Markov model (HMM), Long Short Term Memory's do not need to initially commit to keeping a fixed number of states. Input and output biases are just two of the many adjustable features of Long Short Term

Memory's. Therefore, there is no need for minor adjustments. Like Back Propagation Through Time (BPTT), Long Short Term Memory's reduce the complexity of updating each weight to O (1).

Training a network aims to reduce the amount of error or expense introduced into the final product when the network processes training data. To discover the ideal set of weights with the smallest loss, first calculate the gradient or loss for a given set of weights, then adjust the weights as needed. The concept of backtracking is based on this premise. This is not uncommon for the gradient to be so slight and barely perceptible. This is important to remember that certain factors in the layers that follow can impact the gradient of the layer in question. The resulting gradient will be even smaller if some of these components are small (less than one). This is referred to as the scaling effect. Multiplying this gradient by the learning rate, a small amount between 0.1 and 0.001, reduces the size of the final result. Therefore, the weight shift is negligible, yielding the same results.

Similarly, the weights are adjusted to a higher than optimal value if the gradients have a significant value because of large component values. The problem of extending gradients describes this situation. The extending of gradients or the backward gradient problem, as noted by usually faced in the area of deep learning, is the problem of effective backpropagation in deep learning models. In the case of very deep networks, for example very deep convolution neural networks, the gradients can either decrease to a state, which is called the vanishing gradient problem and may cause challenges in weight updates for shallow layers. This problem limits the application in the training process, resulting in poor convergence or oscillatory weight updates, and therefore mitigation strategies such as gradient clipping or other structures such ramp cut off

recurrent units or residual networks are needed. The neural network unit was reconstructed with the scaling factor set to one, removing the effect of scaling. The cell was later given the moniker Long Short Term Memory and improved with more gating components.

The Long Short Term Memory network consists of 64 cells with recurrent layers and three gates: input, forget, and output. Every Long Short Term Memory gate has a unique function that aids in determining the output for a given input. Furthermore, the cells contain memory units that store all processing information. Memory cells are more valuable in simplifying the process of computing outputs. The memory cell updates the processing inputs in a predefined time in each computation. The network has one recurrent layer, 64 cells, and a dropout layer to discard extraneous information during computation. The dropout layer reduces overfitting issues that cause the generalization and prediction accuracy to be maximized. Then the below mathematical computations are utilized to find the output value for diabetic data.

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{3.21}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{3.22}$$

$$O_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{3.23}$$

$$g_t = \sigma_t\left(W_g x_t + U_g h_{t-1} + b_g\right) \tag{3.24}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_t(g_t) \tag{3.25}$$

$$h_t = O_t \circ \sigma_t(c_t) \tag{3.26}$$

$$y_t = O_t \circ \sigma_t(c_t) + x_t \tag{3.27}$$

In the above computations from (equations 3.21 to 3.27), input sequences are denoted as the $x_t = \{x_1, x_2, \ldots . x_T\}$, output sequence is represented as $y_t = \{y_1, y_2, \ldots . y_T\}$,

W, U, and b are denoted as the network parameters such as weight and bias. Here, c_t is represented as the memory cell vector, forget gate is f_t, input gate is denoted as i_t and output gate is represented as O_t. During the computations, the sigmoid activation function σ_g, hyperbolic tangent σ_t and dot product ∘ is applied to get the output value. Here the computed values are fed into the last recurrent layer that uses the weights and bias values to estimate the net output, which is defined in equation (3.28).

$$Z_i = sigm(\sum_{i=1}^{N} Y_i w_i + b_i) \tag{3.28}$$

In equation (3.28), The fully connected network parameters weights w i and bias value b_i are used to produce the net output Z. The sigmoid activation function is used to calculate the final output, which is as $sigm = \frac{1}{(1+e^{-a})}$. The computed output value is compared to the training dataset to estimate the error value. If the computed values differ from the real value, the network parameters are changed to reduce the misclassification error rate. The flock optimization working behaviour is applied to pick the optimum parameters in the search space. This has been seen that performance and results of any optimization technique can greatly change with respect to any hyper parameter change, such as, learning rate, batch size, number of layers, dropout rate, etc. One's model performance with respect to convergence, overfitting or prediction accuracy against new data may be very much influenced by even small alterations in these numbers. For example, if a learner's rate is too high, the learner would leap to and even overshoot the optimal solution making things unstable while if this is too low, then convergence would be reached very late or might even be stuck at the local optimum. In such conditions, elevation of the batch size, number of epochs or shrinkage of the dropout rate would also lead to poor learning while overfitting would

be observed. Thus, hyperparameter tuning in a model is useful as this ensures that the trade-offs between the model training time, the level of complexity of the constructed model as well as the accuracy of the predictions made are reasonable and proper. The selected optimization algorithm-based technique reduces classification error rates and enhances overall diabetic illness prediction accuracy. According to the above discussion, the algorithmic steps for flock optimized deep neural model is described in Table 3.9.

Table 3.9: Steps for Flock Optimization-based Deep learning algorithm

Initialize: input I ($i_1, i_2, i_3, \ldots i_n$); features F ($f_1, f_2, \ldots f_n$), weight w, bias b, search space S with dimension D, fitness function $f(x)(initial = 0)$, total number of elements in the flock n_b, number of birds in the flock n_{cb}, inertial coefficient ω, cognitive coefficient λ, social coefficient γ, and topological coefficient h, Velocity μ_k and position of the bird x_k
(pre-processing)
Step 1: convoluting I with $G_\sigma(x, y)$ // Gaussian function $G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}$ Step 2: detect the outlier information using Z-score normalization Step 3: replace the missing value using mean filter The output of step 1: filtered input (fI)
(Feature extraction)
Step 3: Analyse S and divide the data according to different characteristics (sex, insulin, glucose level, BMI, age, etc.). For every data, **analyse** (x_k, μ_k) **Evaluate** Fitness $f(x) = f(x_k^{(1)}, x_k^{(2)}, x_k^{(3)}, \ldots x_k^{(D)})$ **Compare** fitness with the particle's best fitness value, then assign the current position as the best position **Update the** best fitness **Compare** fitness with the best global fitness **Update** global fitness **End loop** Transfer fI into multi-layer convolutions to get the feature vector F. // defined convolution as z[m]=$\sum_{i=-l}^{l} x[i].\delta[m-i]$, $x[i]$ Input, kernel - δ ($size$ 8), convolution result is z with index m, in the first layer. // Step 4: Transfer the z[m] into the next convolution layer with the pooling function to obtain the feature map (y^*) which is obtained after performing down-sampling in that convolution layer. Step 5: Then maximum feature map (Y) is obtained by examining (y^*). Step 6: The final Y-predicted convolution layer fed into the recurrent layer
(Prediction)

// here Long Short Term Memory network is trained by using respective inputs and outputs, which helps to predict the incoming inputs pattern//
//The network uses the input vector, output vector, and memory cells to analyse input data//
Step 7: Fed Y into the input, forget, and output gates.
Step 8: Get Y_{out} value from the output gate.
Step 9: Update (w and b) according to the flock optimization process defined in step 2
Step 10: compute the overall out using a sigmoid function (Z_i)
Step 11: Predict the patterns as abnormal or normal.

According to the above algorithmic steps, the input diabetic information is processed, and classification output is obtained. The flock optimizing technique successfully identifies the data availability and the relationship using the sampling process. Then each sample is examined, and sequence attributes are detected according to the flock optimization process. Detecting the attributes and relationships helps resolve the computation complexity because the interrelationship data only identifies the most relevant patterns-related attributes. Further, the optimized weight updating process defined in the prediction process reduces the misclassification rate. Then the introduced flock-optimized deep learning model-based diabetic disease prediction model efficiency is described in chapter 4.

CHAPTER 4

PERFORMANCE ANALYSIS

4.1 Introduction

Diabetic detection is one of the important concepts in the health industry because as this is one of the major diseases exists widely. Almost every person suffers from the diabetic disease, which leads to stroke, heart problems, etc. According to chapter 2, the researchers use machine learning techniques, Data mining approaches, Artificial Intelligence (AI), and optimization models to predict diabetic diseases. These methods sufficiently utilize the respective functions and learning processes to predict diabetic diseases. However, these techniques face difficulties while analysing a large volume of data. In addition, the existing systems consume high computation difficulties during the data association identification process. The traditional systems have complexities and create over-fitting issues, convergence problems, complex prediction, predominant features, and non-convergence optimization issues. These issues are overcome by applying the Flock Optimization Assimilated Deep Learning Model (FOADLM)-diabetic detection process. The algorithm uses the Gaussian filtering approach to eliminate the outlier information during the analysis. Then different layers are utilized to analyse the noise-removed information. The first layer uses the flock optimization algorithm that minimizes difficulties while analysing the correlation between the features in the dataset. The flock uses distance measures to identify the relationship between the data. The predicted distance value identifies the association and movement between the flocks. The feature is extracted from the dataset according

to the flock working process. The derived features are processed using the Long Short Term Memory approach to predict normal and abnormal features. The created FOADLM based created approach's efficiency is evaluated using experimental analysis. Hence this chapter discusses the excellence of the FOADLM-based diabetic prediction process.

4.2 Implementation Settings

The Flock Optimization Assimilated Deep Learning Model (FOADLM) implementation in MATLAB involves the first step of loading the dataset then pre-processing. The dataset is further partitioned to comprise 60% which is used for training and 40% used for testing. This is because there are efficient matrix operations and machine learning toolbox available in MATLAB that are very useful for this implementation. The efficiency of the system is evaluated by splitting the dataset into 60-40% because it is less computation expensive, faster and simple while determining the introduced system efficiency. The amount of noise in the images can be minimized by the procedure of pre-processing whereby Gaussian convolution is involved and detection of outlying data is executed through Z-score normalization.

In the training phase, the training set is subjected to feature learning through the use of the convolutional layers followed by pooling so as to capture important features. These maps then go through a recurrent layer which seeks to learn the time dependency of patterns (LSTM) The performance of the network during this stage was improved by optimizing the weights and biases via a flock based optimization approach whereby the network optimization process is conducted iteratively by reducing the error of the network with respect to the performance indicators' global and local activities.

In this one the obtained predictive model is then validated with the testing set to determine the respective level of prediction. If the supposition is fulfilled, the model is efficient as predicted in the validation process whereby the prediction is accomplished by propagating the test data through the established model, after which the model generates the final prediction of interest via the utilization of a sigmoid function. Then the efficiency of the system is evaluated using different performance metrics which are listed as follows.

4.3 Performance Metrics and Analysis

This work uses the Flock Optimization Assimilated Deep Learning Model (FOADLM) to perform the diabetic detection process. The Diabetic Disease Classification (DDC) is developed using the MATLAB tool. The MATLAB tool consists of GUI interface tools and in-built functions that are used to analyze the data effectively. The implementation tool has various learning functions and training processes to solve scientific and computing problems. This work uses three different datasets, such as the PIMA Indian dataset, the OhioT1DM dataset, and the Azure diabetic dataset, to analyze the system's efficiency. From the collected information, 60% of data is utilized for training the system and 40% of data utilized for testing. The research uses the mean square error rate (MSE), RMSE, precision, recall, accuracy, Matthew correlation coefficient, computation time, reliability, and scalability metrics. The error rate metrics RMSE, MESE used to justify the model attains minimum deviations while identifying the diabetic disease. The Mean Squared Error (MSE) for Nadesh et al. (2020) was derived by estimating the error rates based on the reported accuracy, sensitivity, and specificity values. By converting these classification metrics into a form that reflects

the average squared difference between predicted and actual values, MSE was approximated, enabling a comparison across models. The accuracy metrics are used to determine the classification accuracy which balances the false and true negative and positive. Additionally, computation time, reliability, and scalability are crucial to assess the model's efficiency, robustness, and ability to handle larger datasets, aligning with the study's goal of creating a reliable, high-performance predictive system. The use of multiple performance metrics is effective because the evaluation of the model will be complete. Accuracy describes the total correctness score of the model while precision and recall address the concern of false positive and false negative respectively. For instance, the Matthew Correlation Coefficient (MCC) metric helps in mitigating the performance across the biased/ imbalanced datasets while MSE/RMSE measures the amount of error in predictions, assuring that several dimensions of model development are enhanced without compromising other aspects.

The efficiency of the introduced Flock Optimization Assimilated Deep Learning Model (FOADLM). The excellence of the FOADLM approach is compared with the different existing methods such as Deep Neural Network (DNN) (Nadesh et al., 2020), improved particle swarm optimization optimizing long short-term memory network (IPSO-LSTM) (W. Wang et al., 2020), gravitational search optimized deep learning (GSODL) (Ezzat et al., 2021), tubu optimized sequence modular network (TOSMN) (AlZubi et al., 2020), Adaptive Neuro-Fuzzy Inference System (ANFIS) (F. Haque et al., 2021), Ensemble Learning Approach (ELA) (N. L. Fitriyani et al., 2019), and DiabDeep (H. Yin et al., 2021). The existing methods are discussed in Chapter 3. From the researcher's opinions, the obtained results are discussed to analyze the efficiency of the FOADLM system. The 200 epochs were set to ensure the model has sufficient

iterations to learn complex patterns from the data without overfitting or under fitting. This number strikes a balance between allowing the network to converge while avoiding excessive training that could lead to diminishing returns in performance improvement.

4.3.1 Mean Square Error Rate (MSE)

The Mean Square Error Rate (MSE) is also named the Mean Squared Deviation (MSD). The MSE metric is used to compute the average error value while computing the output value. The error rate is estimated by taking the difference between the actual and estimated values. The computed MSE value is related to the anticipated value of the squared error loss. The computed MSE value is always positive and minimum. The MSE value is estimated using equation (4.1).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{4.1}$$

In equation (4.1), Y_i is defined as observed value and $\hat{Y}_i$ is defined as the predicted output value. According to the computation, the obtained MSE results are illustrated in Figure 4.1.

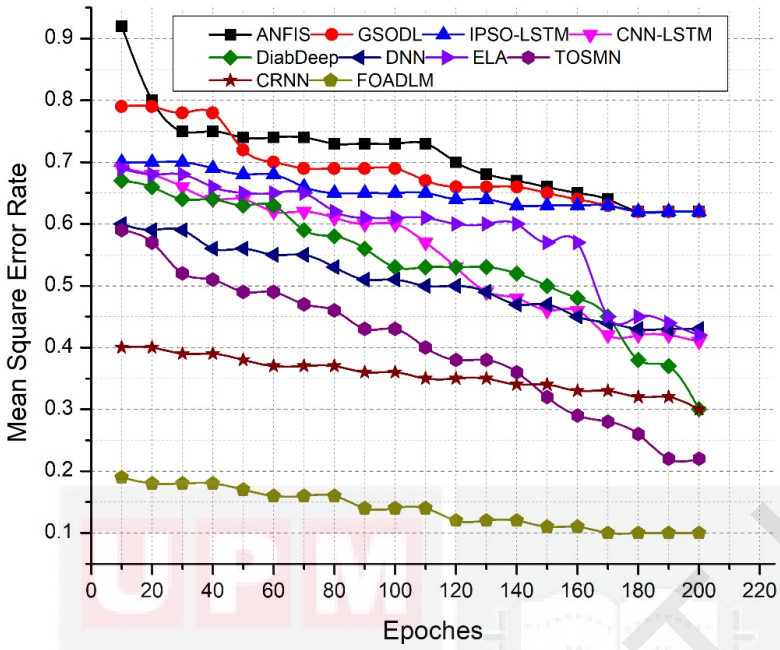


Figure 4.1: Mean Square Error Rate Analysis

Figure 4.1 demonstrates the mean square error rate analysis of the introduced Flock Optimization Assimilated Deep Learning Model (FOADLM) based diabetic prediction process. The results show that the introduced FOADLM approach has a minimum error rate value compared to the other methods. The efficiency is examined for different numbers of epochs, and the system has a minimum error value. The introduced approach has the gaussian filter that removes noise from the inputs without creating difficulties. The filter uses the kernel values that are convoluted with the input data to replace the irrelevant and inconsistent information. The effective utilization of the Gaussian filtering process eliminates the irrelevant and inconsistent details that maximize the overall diabetic prediction rate. In addition, the flock starling bird's behaviour is utilized to analyse the relationship between the data. The data associations are computed in the search space that identifies similar data. According to the similarity, the network predicts the feature map. The computed feature values are more useful for recognizing the diabetic features in the search space. The selected features

are processed by the Long Short Term Memory network, which has a memory unit that stores the processing inputs. The successive training process detects the sequence of inputs with minimum computation complexity and deviation between the output values (Figure 4.1).

4.3.2 Root Mean Square Error (RMSE) Rate

RMSE is also called Root Mean Square Deviation (RMSD). The root-mean-squared error is calculated as the difference between the estimated and forecasted output values. The RMSE value is estimated from the MSE value, which means the square root of the sample moment of deviations. The deviations are generally called the residual value, which is computed using equation (4.2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{4.2}$$

The above equation (4.2) is used to identify the Root Mean Square Error Rate (RMSE) of the introduced Flock Optimization Assimilated Deep Learning Model (FOADLM)-based diabetic classification system. According to the computation, the obtained output values are illustrated in Figure 4.2.

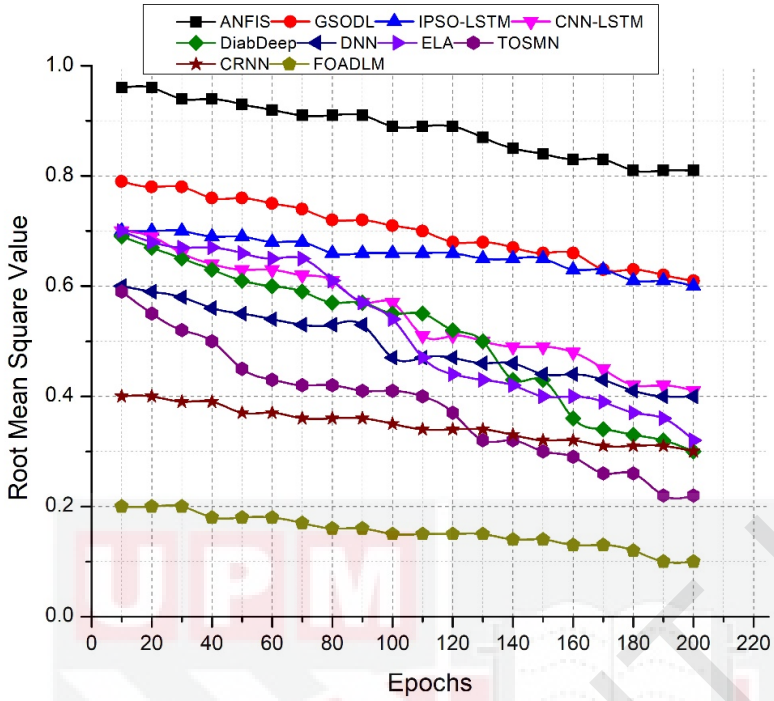


Figure 4.2: Root Mean Square Error Rate (RMSE) Analysis

Figure 4.2 demonstrates the graphical analysis of the RMSE Rate analysis of the Flock Optimization Assimilated Deep Learning Model (FOADLM). The FOADLM approach efficiency is compared to the existing methods defined in chapter 3. The results are analysed for different epochs in which the FOADLM approach attains minimum RMSE value. The flock optimization algorithm uses the fitness function $f(x) = f(x_k^{(1)}, x_k^{(2)}, x_k^{(3)}, \dots x_k^{(D)})$ to select the feature map in the search space. The selected features are analysed using the multi-layer convolution network as $z[m]=\sum_{i=-l}^{l} x[i].\delta[m-i]$ that predicts every feature involved in the classification process. In addition, flock starling bird's movement and distance measures are utilized to select the network parameters. The selected parameters are more useful for updating the network performance. The frequent updating process minimizes the overall deviation between the computed output values. The minimum deviation indicates the Flock Optimization Assimilated Deep Learning Model (FOADLM) attains high prediction efficiency.

4.3.3 Precision

Precision is named the Positive Predictive Value. This metric measures how effectively the introduced approach identifies the relevant samples from the retrieved samples in the search space. The precision is computed using equation (4.3).

$$Precision = \frac{True\ Positive}{True\ positive + false\ positive} * 100 \quad (4.3)$$

In the above equation (4.3), TP is referred to as the True positive means correctly classified diabetic data, TN is defined as true negative is correctly identified healthy people, FP is defined as false positive means, incorrectly identifies diabetic people, and FN is false negative means incorrectly identify the correct feature.

4.3.4 Recall

The recall is named the Sensitivity. This metric measures how effectively the introduced approach retrieves the relevant instances. The precision is computed using equation (4.4).

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} * 100 \quad (4.4)$$

According to the above discussion, the Flock Optimization Assimilated Deep Learning Model (FOADLM) efficiency is evaluated with existing methodologies. The obtained precision-recall related graphical analysis is illustrated in Figure 4.3.

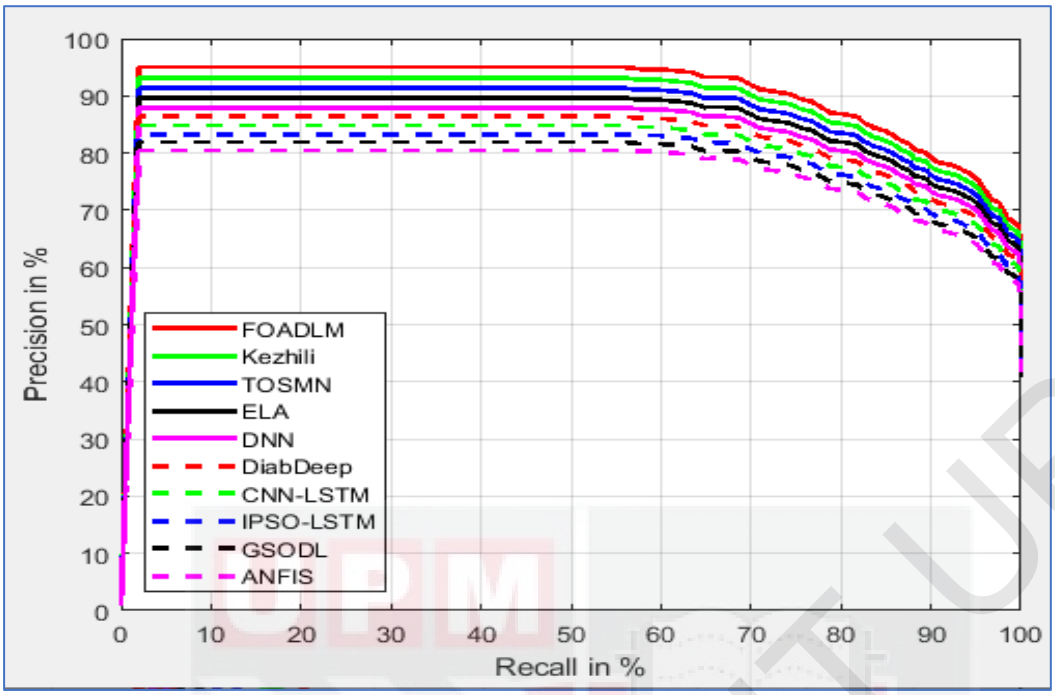


Figure 4.3: Precision and Recall Analysis

Figure 4.3 illustrates that the Flock Optimization Assimilated Deep Learning Model (FOADLM) successfully retrieves the diabetic feature relevant information from the search space. The flock optimization model applied on the convolution neural networks layer one that effectively utilizes the flocks association. The association between the data has been determined according to the distance measure. The computed distance value and radius information predict the feature space's attractive and collision-related features. The computed features are arranged to generate the feature map in the search space. The feature map is analysed using the convolution layer that predicts the diabetic disease-related features. According to the figure, the FOADLM approach attains a high precision value. The Long Short Term Memory network sequentially processes the input, and the selected features are then examined in the search space. The sigmoid activation function is utilized during the analysis to compute the output value. The network uses the flock optimization model to update the network parameters such as weight and a bias value. The frequent network

updating procedure minimizes the deviation between the outputs. The minimum deviation value indicates the FOADLM approach attains the maximum recall value compared to the other methods.

4.3.5 Accuracy

Accuracy is the metric used to compute the introduced approach's recognition efficiency. The accuracy is computed using equation (4.5).

$$\frac{Number\ of\ instances\ classifed\ correctly}{total\ number\ of\ instances} * 100 \tag{4.5}$$

Equation (5.5) illustrates the computation of accuracy metrics while classifying the diabetic features. According to the above computation, the obtained result is shown in Figure 4.4.

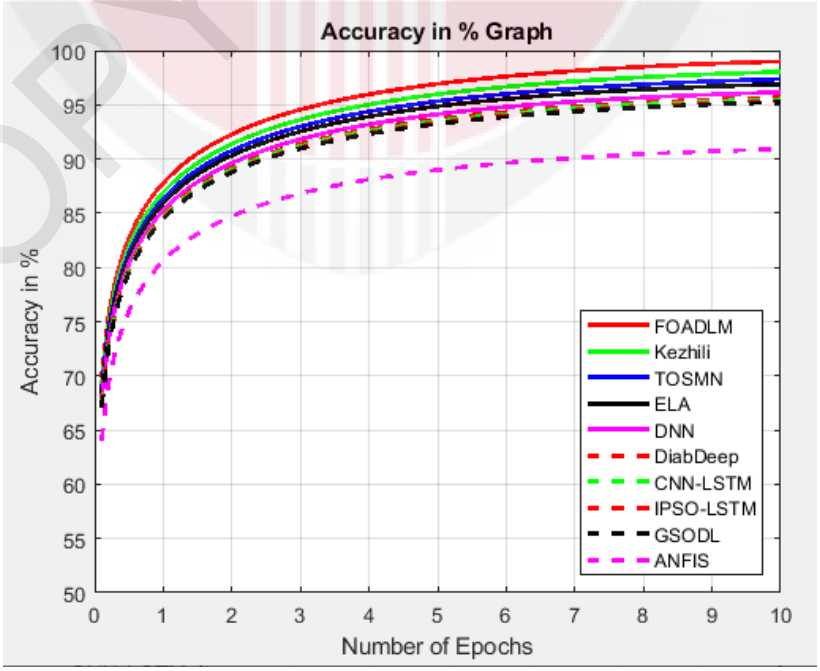


Figure 4.4: Accuracy Analysis

The Flock Optimization Assimilated Deep Learning Model (FOADLM) successfully extracts every feature relevant information from the search space, as shown in Figure 4.4. The flock optimization model was used for the convolution neural networks layer one, which successfully uses flocks association. The distance measure was used to determine the relationship between the data. The estimated distance value and radius information are utilized to predict the attractive and collision-related characteristics of the feature space. The computed features are organized in the search space to form the feature map. The convolution layer predicts diabetic disease-related features and is used to assess the feature map. According to the graph, the FOADLM method achieves a high precision value. The selected characteristics are subsequently evaluated in the search space, which employs the Long Short Term Memory network to successively process the input. The sigmoid activation function is employed during the analysis to calculate the final value. The flock optimization model updates network parameters like weight and bias value. The frequent network update technique reduces the difference in results. Since the FOADLM strategy has the smallest deviation which must have the best accuracy value among the other approaches (described in Chapter 3). The results of accuracy based on the number of epochs for comparative models were obtained by tracking the models' performance after each epoch during training. The accuracy at different epochs was recorded to observe how learning progressed and to compare convergence rates across models, ensuring a fair evaluation of the predictive capabilities over time.

4.3.6 Matthew Correlation coefficient (MCC)

Mathew Correlation Coefficient (MCC) is the measure used to compute the efficiency and quality of the introduced approach while making the classification. The MCC coefficient uses the True Positive, False Positive, and False Negative to estimate the system's efficiency. The MCC has a value between -1 to +1. The MCC is computed using equation (4.6).

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4.6)$$

According to equation (5.6), the MCC value for Flock Optimization Assimilated Deep Learning Model (FOADLM) is computed, and the graphical analysis of the MCC is illustrated in Figure 4.5.

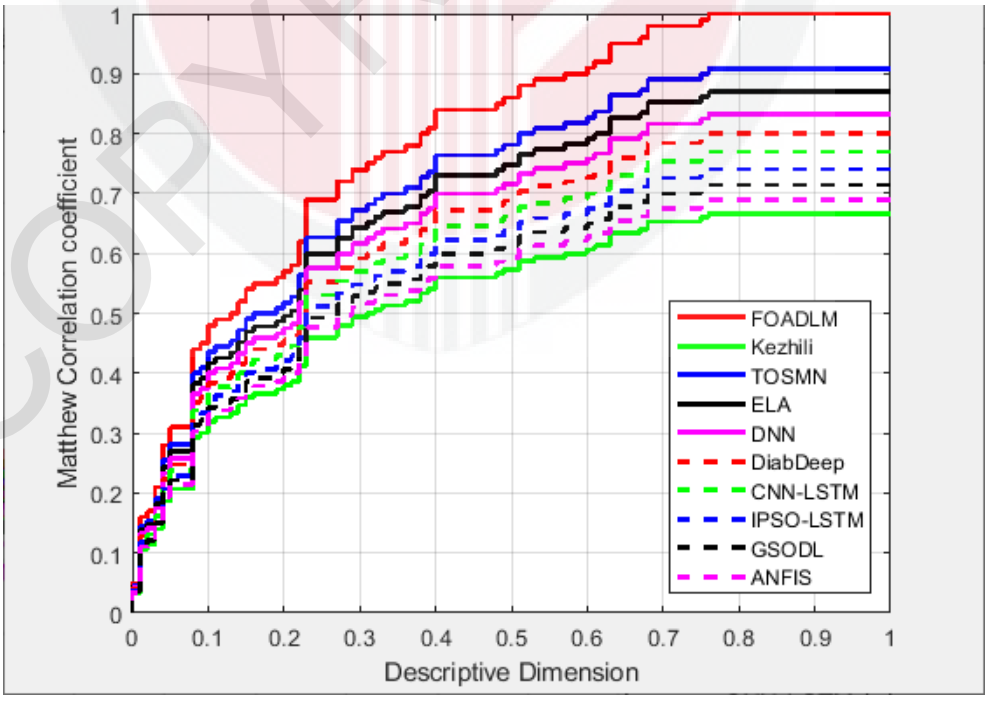


Figure 4.5: Matthew Correlation Coefficient Analysis

Compared to previous approaches, the FOADLM methodology achieves the lowest error value. The FOADLM approach optimizes network parameter selection by regularizing and optimizing network parameters. Compared to previous methodologies, effective backpropagation of values reduces the variance between the actual and computed values. The sampling technique, in this case, use probability values to estimate related features in the feature list, which is employed to simplify the sequence identification process. Furthermore, the flock optimization algorithm leverages the bird positions described in chapter 4 to aid the search process. Agent attributes are modified for each search to reduce calculation challenges. The bird-searching strategy is used in the neural network parameter updating process to minimize differences between the real and predicted values. The FOADLM achieves the lowest RMSE value, indicating that the computed output values effectively match the anticipated line point. The computed data points are compared with the training patterns, which aids in determining the difference between the actual and anticipated values. The method employs flock behavior to determine network parameters, lowering error and enhancing reasonable data rate. The low mistake rate implies that the system achieves good recognition accuracy (Figure 4.5). The FOADLM method has a high level of precision. The high precision rating implies that the system correctly predicts diabetes data. In this case, the deep learning network employs flock optimization techniques in the network layer to properly assess the feature. The derived features efficiently address the issue of data availability. Recurrent neural networks employ gates that combine network weight and bias (w and b) values with an activation function. The network employs memory cells, which continuously store and update previously processed data. To forecast the output $y_t = O_t \circ \sigma_t(c_t) + x_t$, the memory cell, input, output, and forget gates ($c_t = f_t \circ c_{t-1} + i_t \circ \sigma_t(g_t)$) are more

beneficial. Appropriately using the activation function and network parameters accurately identifies normal and diabetic features. However, the strategy minimizes computing time, maximizing reliability and scalability while assessing diabetes characteristics. The collected results are then depicted in Figure 4.6.

4.3.7 Computation Time

The computation Time metric is used to predict the time value for the introduced approach to recognize the diabetic features from the set of features in the search space.

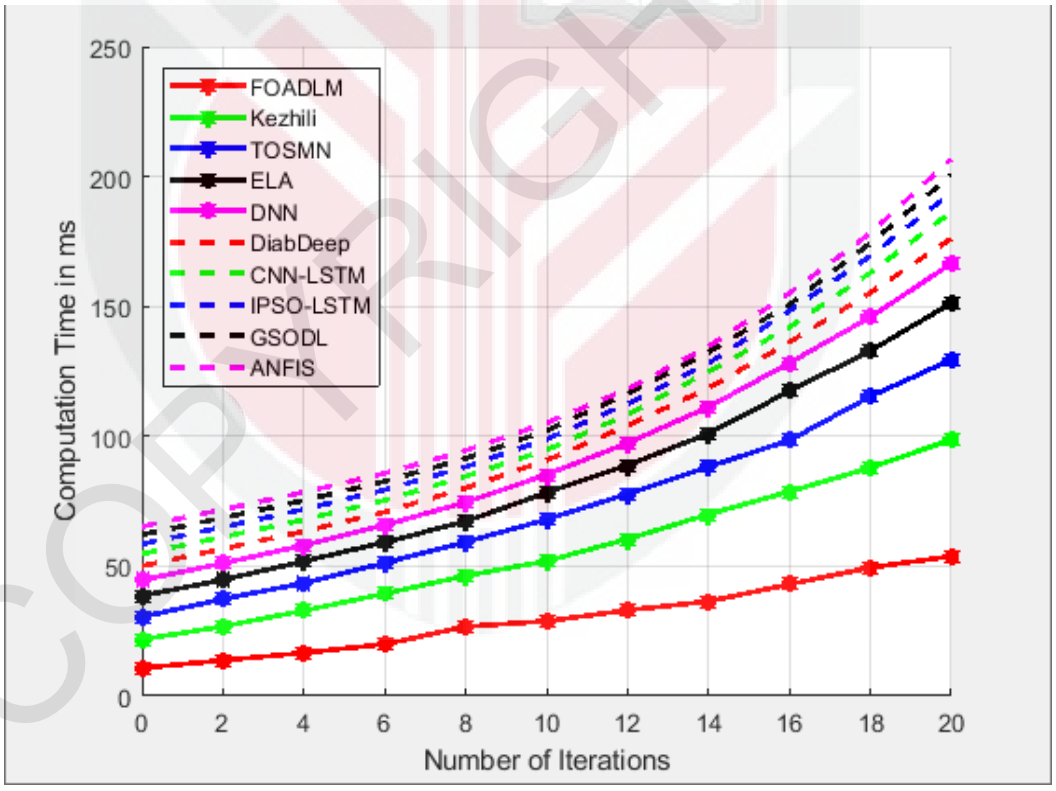


Figure 4.6: Computation Time Analysis on PIMA Indian Dataset

Figure 4.6 demonstrates the computation Time analysis of the FOADLM-based diabetic disease classification system. The results use the PIMA Indian diabetic

dataset, and in 20 iterations, the performance is carried out. The results clearly state that the introduced FOADLM approach requires minimum computation time while processing the features. Here, the convolution neural model and Long-Short Term Memory Network are utilized to analyze the input diabetic data. The convolution network uses the flock optimization and sampling technique to normalize the data while processing the input. The convolution model incorporated the flock optimization technique that identifies the relationship between the data. The identified relationship value utilized in the training set minimizes the difficulties while classifying the data in the search space. In addition, the Long Short Term Memory network has memory units that store every processing input, which minimizes the computation time while analyzing the testing data efficiency. Thus the FOADLM approach attains a low computation time for the different number of data and iterations.

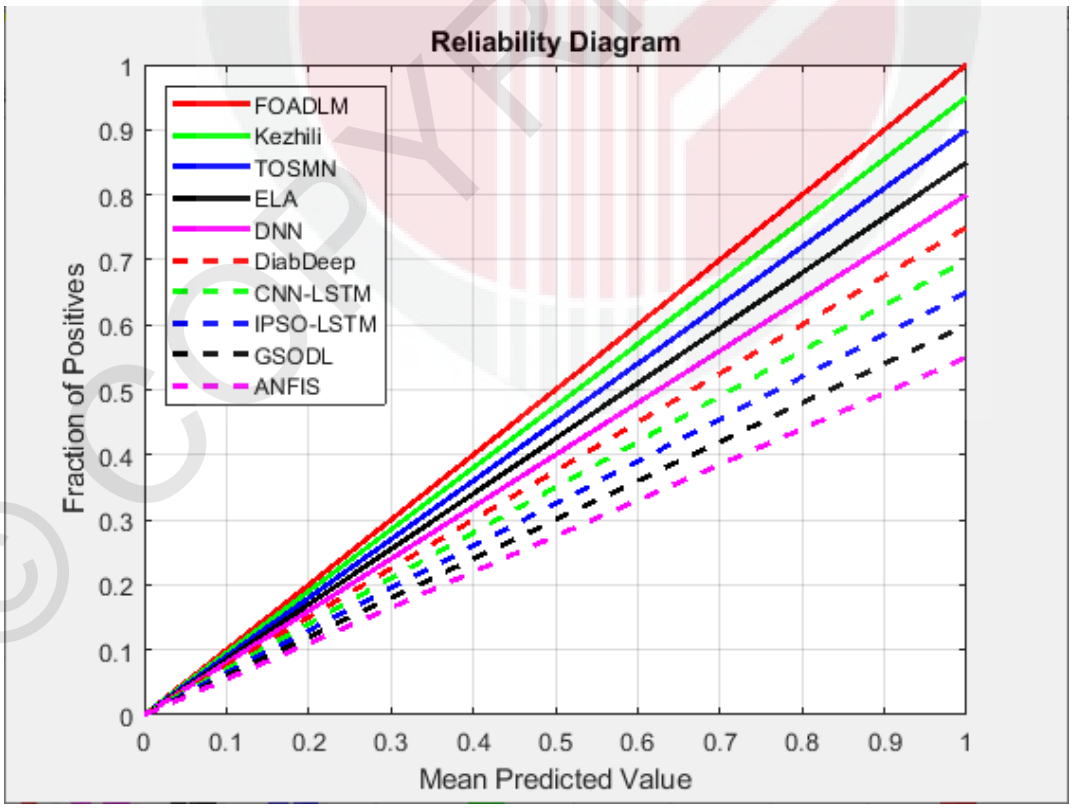


Figure 4.7: Reliability Analysis

Reliability is the measure used to identify the consistency and trustworthiness of the FOADLM approach while classifying diabetic data. Figure 4.7 illustrates the reliability analysis of the FOADLM approach compared with the several existing approaches defined in chapter 3. The results are computed between the mean predicted value and the fraction of positive values. The results clearly state that the FOADLM approach predicts the correct samples and eliminates irrelevant or false information in the search space. The algorithm uses block initialization, sequence identification, distance computation, and association identification procedures. The result of these methods is highly predictive of the input. The system's accuracy is increased by comparing the anticipated output value to the training set. In addition, the Long Short Term Memory model analyzes the sequence of inputs with predefined functions and learning parameters. The learning process uses the memory cell and different gate values to identify the exact output by managing the system reliability measure.

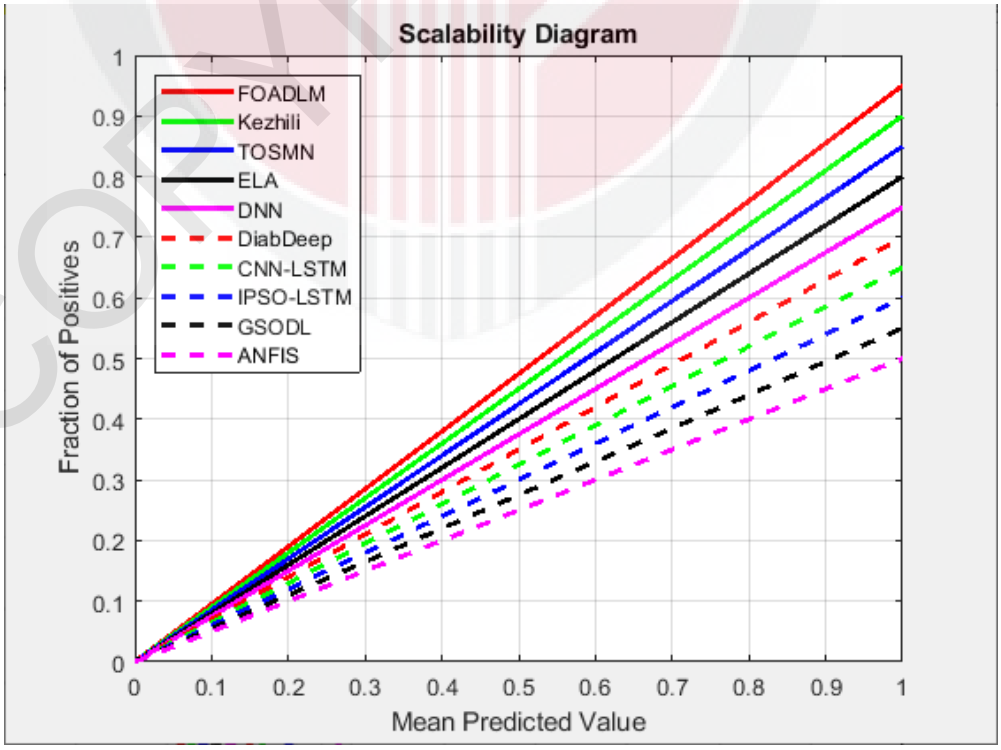


Figure 4.8: Scalability Analysis

An easily extensible system may adapt to varying workloads and user requirements. The ability of a system to scale up or down in response to fluctuating demands is known as scalability. According to figure 4.8, the system scalability is analyzed, and the system's efficiency is compared with existing methods. The results clearly show that the introduced FOADLM approach manages the system's scalability and processes the huge volume of information. The approach uses the Simple Random Sampling (SRS) algorithm to analyze every piece of information. The sampling technique uses cluster details, which helps predict the relationship between the data. The sampling technique group similar information that reduces the computation difficulties. In addition, the Long Short Term Memory network predicts the training template for each input using the memory unit that contains the whole processing input. The effective utilization of the training, activation function, and learning procedure helps to manage the system's scalability.

Effectively using each function in the network improves the overall recognition rate. Furthermore, the method includes a flock optimization procedure that minimizes the misclassification error rate. In addition, the method uses the learning parameter and objective functions that reduce the data availability issues in various rounds. The continuous training method increases the overall prediction rate and classifies the features with the least calculation time. During implementation, sampling approaches are used to study the data properties and relationships, reducing variations in computed output.

Furthermore, this promotes system flexibility, reliability, and scalability. This efficiency is measured using MSE and RMSE measures. The Long Short Term

Memory approach is used in researcher study to train the data feature, which helps to shorten the computation time of diabetic prediction. Furthermore, this correlates data from the previous step of the analysis. This training procedure enhances the overall connection between the data assessed using the Matthew correlation coefficient and precision measures. The entire efficiency attained is then shown in Table 4.1.

Table 4.1: Overall Accuracy Analysis

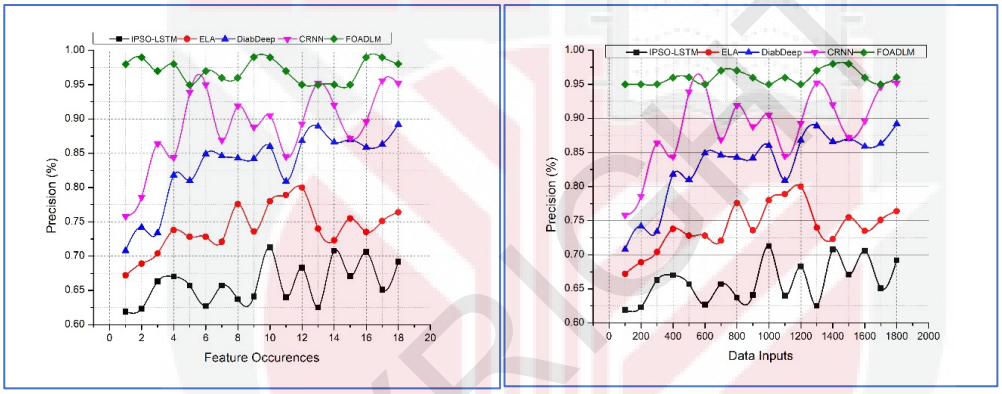
S.No	Methods	Accuracy
1	Deep Neural Network (DNN)	96.16%
2	Improved Particle swarm optimization optimizing Long short-term memory network (IPSO-LSTM)	95%
3	Gravitational search-optimized deep learning (GSODL)	95%
4	Tubu optimized sequence modular network (TOSMN)	98.52%
5	Adaptive Neuro-Fuzzy Inference System (ANFIS)	91.17%
6	Ensemble Learning Approach (ELA)	96.74%
7	DiabDeep	95.3%
8	CRNN	98.63%
9	FOADLM	99.23%

Table 4.1 clearly shows that Flock optimization-based deep learning model attains a high accuracy (99.23%) compared to the other method, such as Deep Neural Network (DNN) (96.16%) (Nadesh et al., 2020), improved Particle swarm optimization optimizing Long short-term memory network (IPSO-LSTM) (95%) (W. Wang et al., 2020), gravitational search optimized deep learning (GSODL) (95%) (Ezzat et al., 2021), tubu optimized sequence modular network (TOSMN) (98.52%) (AlZubi et al., 2020), Adaptive Neuro-Fuzzy Inference System (ANFIS) (F. Haque et al., 2021) (91,17%), Ensemble Learning Approach (ELA) (96.74%) (N. L. Fitriyani et al., 2019), DiabDeep (H. Yin et al. 2021) (95.3%) and Convolution Recurrent Neural Networks (CRNN) system (98.63%).

Based on these findings, compare the FOADLM method's performance on the PIMA Indian Dataset, the Azure Dataset, and the OhioT1DM Dataset. During the analysis,

the FOADLM approach efficiency is compared with Improved Particle Swarm Optimization optimizing Long Short-Term Memory network (IPSO-LSTM), Ensemble Learning Approach (ELA), DiabDeep, and Convolution Recurrent Neural Networks (CRNN). The comparison methods are selected from the above analysis in which the existing methods attains reasonable efficiency collated with the other methods on various datasets.

4.4 Efficiency Analysis on PIMA Indian Dataset

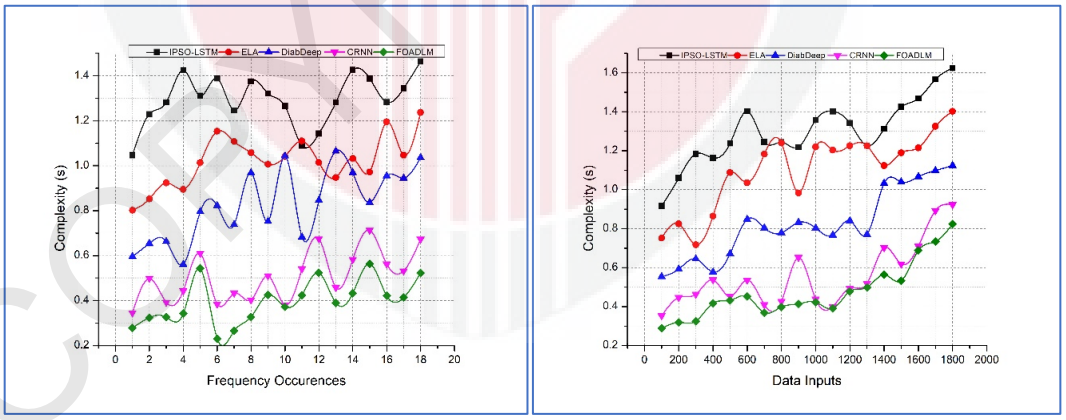


(a) (b)

Figure 4.9: Precision Analysis for (a) Frequency Occurrence and (b) Data Inputs on PIMA Dataset

Across all scenarios of frequency occurrences and data inputs, the Flock Optimization Assimilated Deep Learning Model (FOADLM) shows amazing precision when processing the PIMA Indian Diabetes dataset (Figure 4.9). In order to better understand the resilience and dependability of FOADLM, conduct a precision analysis by evaluating its performance in a variety of data settings. The precision of FOADLM is the result of a number of factors working together. To start, this uses powerful feature selection algorithms to zero in on the most important diabetes-related variables, hence improving the accuracy with which detects genuine positives. Flock optimization

is used to fine-tune a model's parameters and boost its predictive power. To further ensure great accuracy, FOADLM undergoes thorough hyperparameter adjustment to accommodate different data situations. Accuracy can be preserved in the face of fluctuating data occurrences or inputs by employing adaptive decision thresholds based on the properties of the dataset. By incorporating regularization approaches to avoid overfitting and keep predictions stable, FOADLM is able to address class imbalance issues through data balancing strategies like oversampling or undersampling. Its flexible design can process a wide variety of inputs with equal efficiency, and domain-specific information improves its accuracy by directing feature selection and prediction. Overall, FOADLM's capacity to continuously attain high precision across multiple real-world data sets places as strong and trustworthy tool for diabetes identification, with potential applications in a wide range of healthcare sceneries.



(a) **(b)**

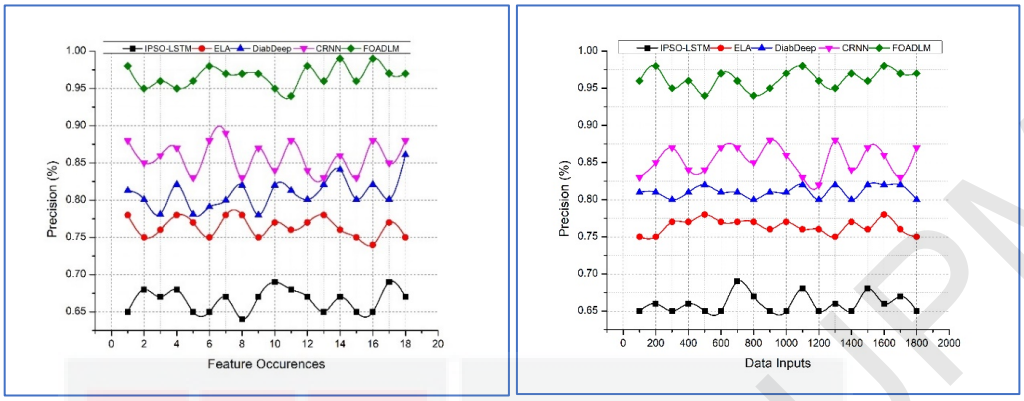
Figure 4.10: Complexity Analysis for (a) Frequency Occurrence and (b) Data Inputs on PIMA dataset

Conducting a complexity analysis of the Flock Optimization Assimilated Deep Learning Model (FOADLM) reveals its remarkable ability to maintain low complexity while processing the PIMA Indian Diabetes dataset under diverse conditions of

frequency occurrences and data inputs (Figure 4.10). In this complex analysis, FOADLM is evaluated across a spectrum of data variability, encompassing different frequencies and data inputs to assess its adaptability and efficiency. FOADLM attains low complexity through several strategic mechanisms and efficiently selects relevant features, reducing data dimensionality, thus model complexity. The model architecture itself is optimized, striking a balance between model size and predictive power. The flock optimization algorithm, a critical component of FOADLM, streamlines parameter optimization, ensuring high accuracy without inflating complexity.

Hyperparameter optimization fine-tunes the model, maintaining a favorable complexity-performance trade-off across various data scenarios. Model pruning techniques remove redundancies, further compacting FOADLM. Adaptive computation adjusts the model's complexity based on the specific data input and frequency occurrences, ensuring efficient resource allocation. Regularization techniques control complexity by discouraging overfitting. FOADLM's ability to consistently achieve low complexity while delivering reliable performance positions as versatile and resource-efficient tool for diabetic detection, adaptable to the intricacies of real-world healthcare data. This balance between precision and complexity ensures its applicability in diverse clinical settings and reinforces its value in healthcare decision-making.

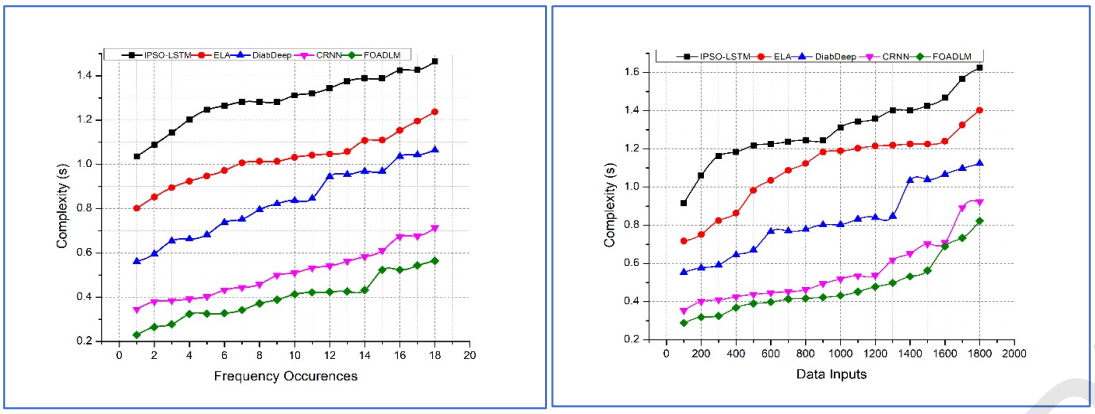
4.5 Efficiency Analysis on Ohio Dataset



(a) (b)

Figure 4.11: Precision Analysis for (a) Frequency Occurrence and (b) Data Inputs on Ohio Dataset

Figure 4.11 illustrated that the precision analysis of the FOADLM appraoch on various number of frequency occurences and data inputs. The analysis uses the Ohio dataset information to analyze the diabetic disease. The ohio dataset is applied to the preprocessing step which is normalized with the help of the z-score normalization that simplifies the processing complexity and improves the overall diabetic prediction rate. In addition, the dataset information is further processed with the help of gaussian filtering and mean filter appraoch that removes the irrelevant data and missing values that also maximzie the overall prediction accuracy. During the analysis, ohio information is processed by assimilating the deep leanring and flock optimization algorithm that identifies the association between the features. The feature association and relationship process predict the diabetic features which helps to improve the overall diabetic prediction accuracy. similarly, the recurrent neural model called Long short term memory network classifies the sequence of predicted features that maximize the overall diabetic prediction accuracy with minimum computation complexity. Then the obtained complexity related graphical analysis illustrated in Figure 4.12.

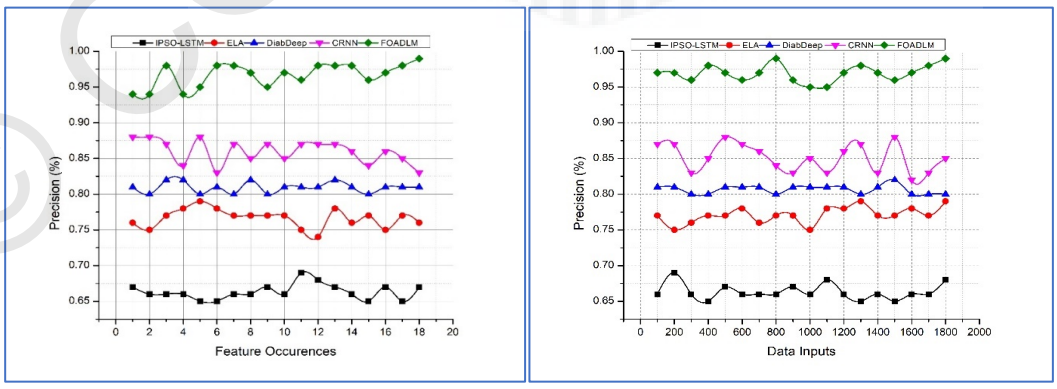


(a) (b)

Figure 4.12: Complexity Analysis for (a) Frequency Occurrence and (b) Data Inputs on Ohio Dataset

From the Figure 4.12, the FOADLM appraoch attains minimum complexity while anlayzing the different number features and input data. The normalization process completely analyze the input features, irrelevant information and missing values that reduce the complexities in the data classification process. In addition, the deep learning process fine-tunes the parameters frequently by comparing the outputs with the template that minimize the error rate. Therefore, the diabetic predicition system minimize the further computations and increases the overall detection accuracy.

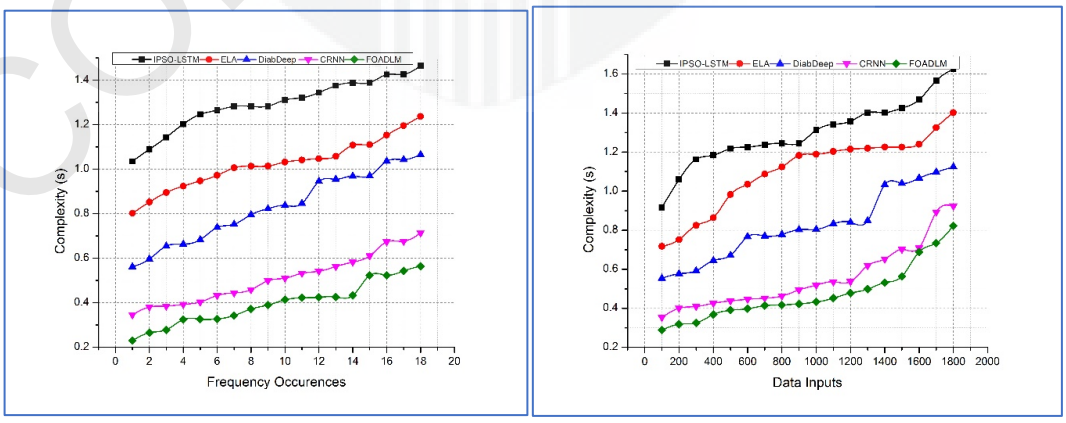
4.6 Efficiency Analysis on Azure Dataset



(a) (b)

Figure 4.13: Precision Analysis on Various datasets for feature occurences and data inputs on Azure dataset

FOADLM, known for its adaptability and precision, can be highly effective in leveraging healthcare datasets like those available on Azure. FOADLM begins by processing, analyzing various patient-related features, including medical history, vital signs, and lifestyle factors. FOADLM excels at feature selection, identifying the most pertinent predictors for accurate diabetic prediction. Its flock optimization algorithm optimizes model parameters efficiently, ensuring that the model is well-suited to the dataset's characteristics. Through extensive hyperparameter tuning, FOADLM adapts to the nuances of the Azure diabetic dataset, fine-tuning its architecture for optimal performance. FOADLM's adaptability shines by handling variations in data, accommodating different patient profiles within the Azure dataset. This maintains a favorable balance between complexity and performance, effectively addressing the challenges posed by real-world healthcare data. This adaptability, combined with its robustness and ability to maintain high precision, positions FOADLM as a valuable tool for diabetic prediction tasks, providing insights that can inform clinical decisions and improve patient care in the context of the Azure diabetic dataset (Figure 4.13). In addition, the FOADLM appraoch attains the minimum complexity analysis compared to other methods.



(a) (b)

Figure 4.14: Complexity Analysis for (a) Frequency Occurrence and (b) Data Inputs on Azure Dataset

4.7 Validation with Existing Kezhili Model

In this section the introduced Flock Optimization Assimilated Deep Learning Model (FOADLM) based diabetic detection System is compared with the existing Kezhili model. The efficiency of the FOADLM approach is compared with the Kezhili model with different metrics such as recall, precision, accuracy and other metrics. The comparative analysis uses the Kezhili datasets for exploring the efficiency; so that, the FOADLM model uses the meal size, insulin dosage, glucose level and exercise variations as input patterns to predict the diabetic disease. The study uses the silico and clinical data which is fed into the FOADLM working process to predict the diabetic disease. The silico data has 180 days of time-series information of 10 virtual adolescent and 10 virtual adult information. Every data has 288 data points that captured patient glucose levels with different variations like insulin administration, time and meal size. The clinical data collected from ABC4D project and ohioT1DM dataset that consists of exercise information, meal details, continuous glucose level and insulin doses. The collected silico dataset information is spilt into 50-50 for testing and training. Similarly clinical dataset also divided into 50% training and 50% testing data.

FOADLM Parameter Setting

The FOADLM is created by initializing the parameter for exploring the working efficiency of diabetic prediction. The network uses the meal/exercise variations, insulin dosage and glucose data as input. Then gaussian filtering is applied for reducing the noise in the dataset. Followed by z-score normalization is applied for normalize the data to identify the outliers. The preprocessed data is fed into the multi-layer convolution networks to captures the latent and predominate features relevant to the

meal-timings, glucoses variability and insulin level. During this process, flock optimization approach is applied for fine-tune the network parameters that minimize the error rate during the training. Finally, the recurrent layer utilizes the Long-Short Term Memory Network (LSTM) to obtain the temporal dependencies for predicting the glucose level. Then the FOADLM system's efficiency is examined using different metrics such as accuracy, MSE, RMSE, precision, recall, computation time, reliability and scalability. Then the obtained results are illustrated in Table 4.2.

Table 4.2: Efficiency Analysis Between Kezhili and FOADLM

Metric	FOADLM (In Silico)	Kezhili (In Silico)	FOADLM (Clinical)	Kezhili (Clinical)
Accuracy (%)	99.23	92.8	99.18	91.6
MSE	0.015	0.025	0.017	0.028
RMSE	0.122	0.158	0.131	0.167
Precision (%)	92.3	90.1	91.4	88.9
Recall (%)	93.7	91	92.5	89.5
Computation Time	18 mins	25 mins	22 mins	28 mins
Scalability	High	Moderate	High	Moderate

Table 4.2 illustrated that the FOADLM ensures the high accuracy value 99.23% compared to the Kezhili systems. In particular, FOADLM quantitatively pronounced its superiority over the other models with respect to both in silico (99.23 % vs. 92.8 %) as well as clinical datasets (99.18 % vs. 91.6 %) thereby showing its enhanced capacity to encapsulate different data types. This surprised with lower Mean Squared Error (MSE) and Root Mean Squared error (RMSE) which implied higher accuracy in the prediction of glucose levels as well as better management of time series fluctuations. Additionally, FOADLM beat Kezhili's model within the accuracy and recall metrics and, in particular, minimized false positive and negative glucose event detections. Further, this emerged as the model used flock optimization which meant that convergence was faster hence less computing time was needed, and flexibility made more easier to apply this for a large data sets and actual clinical situations.

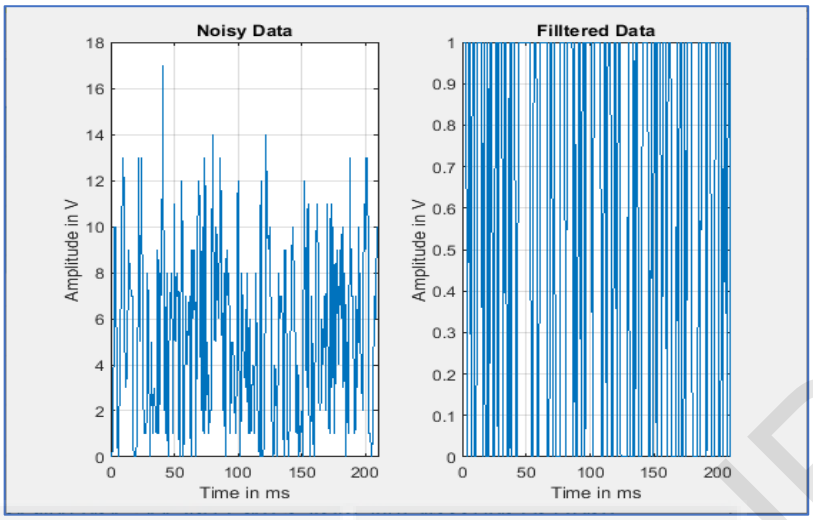
4.8 Results and Findings

This section discusses the efficiency of the Flock Optimization Assimilated Deep Learning Model (FOADLM) while classifying diabetic disease. The FOADLM approach intends to resolve the data assimilation, imbalance, accuracy, and Valuable information loss during the data analysis process.

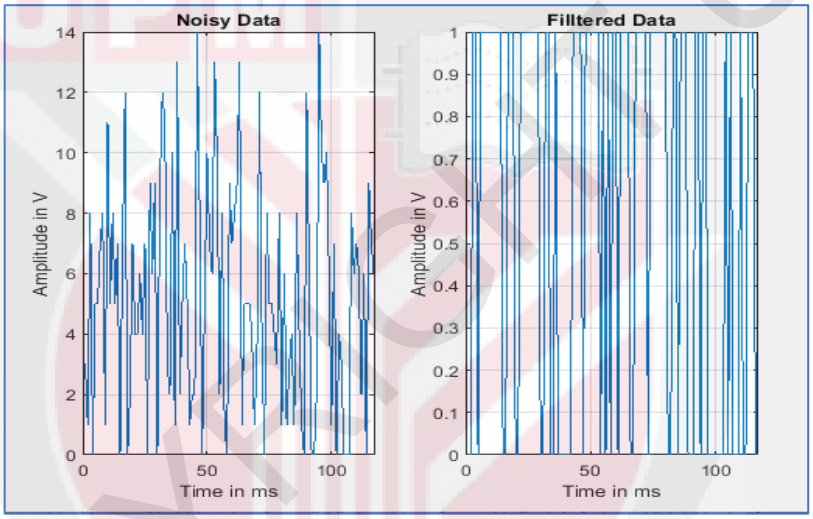
4.8.1 Self-Analysis

Objective 1: To analyse the system flexibility using data sampling techniques incorporated in the first layer as data assimilation incorporated with Flock optimization in CNN

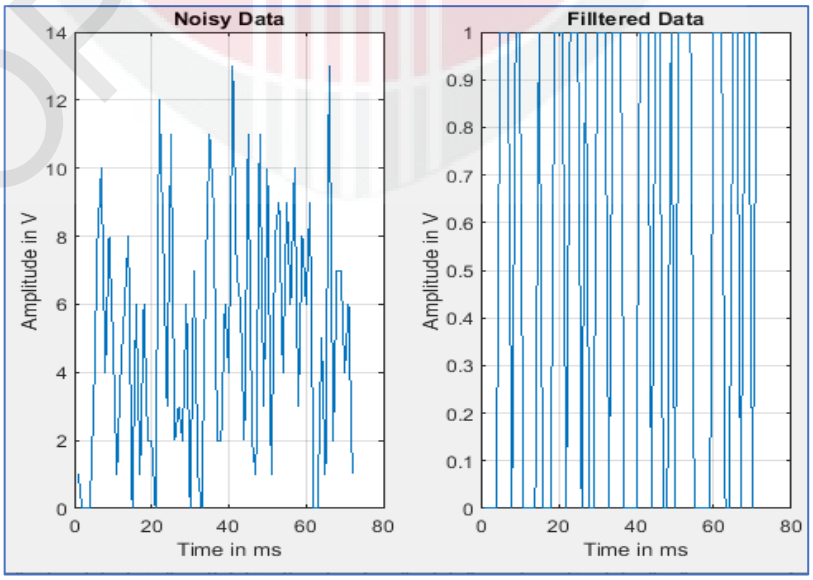
Data assimilation is observing the data and predicting the changes in every data because this affects the data classification process. The collected data is processed by applying the optimization model and theory that identifies the changes in every data. The network predicts the output value based on a variety of factors and training model functions. A neural model that detects patterns in the data is used to process the final result of the computation. In addition, the data having imbalanced data influences the data prediction process and reduces the overall data classification process. Therefore, the main intention of this study is to solve the data assimilation and imbalanced data using noise removal and optimization algorithm involvement. The system's efficiency is evaluated using the three datasets such as PIMA Indian Dataset, Azure Dataset, and OhioT1DM Dataset. The datasets are applied to the MATLAB tool, and the distribution of the data, noise, and filtered data graphical representation is illustrated in Figure 4.15.



(a) PIMA Indian Dataset



(b) OhioT1DM Dataset



(c) Azsure Dataset

Figure 4.15: Noise and Filter Data for different datasets

Figures 4.15 (a), (b), and (c) illustrate the graphical representation of the noise and filtered data of three datasets such as PIMA Indian Dataset, OhioT1DM Dataset, and Azure Dataset. The graphical analysis is demonstrated for time and amplitude, which are used to identify the data distribution. According to the distribution, Gaussian filtering is applied to the kernel value data that is used to eliminate the noise information. After irrelevant and inconsistent information has been weeded out, the threshold value is applied to each piece of data. Effectively reducing global data assimilation problems and data imbalance with this method. Complex, high-dimensional datasets can also reduce filtering efficiency, requiring more advanced techniques to maintain accuracy. More ever, the system uses the flock optimization model that predicts the data relationship, which improves the overall data classification efficiency. After removing the noise from the information, the data was processed by a deep learning model to extract the feature map from the dataset. Among the various researcher's opinions, PIMA Indian Dataset is widely utilized. Hence further analysis uses the PIMA Indian dataset to compute the system efficiency.

Objective 2: To design the diabetic prediction model by investigating the latent and pre-dominant features using the sampling and flock optimization techniques.

The next goal of this effort is to raise the quality of data classification in diabetes. Therefore, the accuracy-related issues are solved by applying the optimized technique in a deep learning network. The learning model has different layers which are processing the inputs using different activation and learning functions to process the inputs. The first layer uses the flock starling bird's activities to analyse the relationship between the data. Moreover, the inputs are handled by max-pooling, convolution, and fully linked layers. The fully connected layer applies weights to the inputs and a bias

value during the output computation. During the analysis, the flock optimization method utilizes the distance measure that predicts the neighbouring data in the search space. The computed distance value chooses the more relevant network parameters. The selected parameters minimize the overall miss classification error rate and improve the overall accuracy. During the implementations, the system uses the Levenberg-Marquardt training function for 1000 iterations. In addition, the network uses ten hidden layers, each with a weight and bias value. As said, the flock optimization algorithm uses the flocking starling bird's behaviour and analyses the data distribution. The flock optimization-based parameter analysis process performed up to 450 iterations, and the parameter searching process efficiency is illustrated in Figure 4.16.

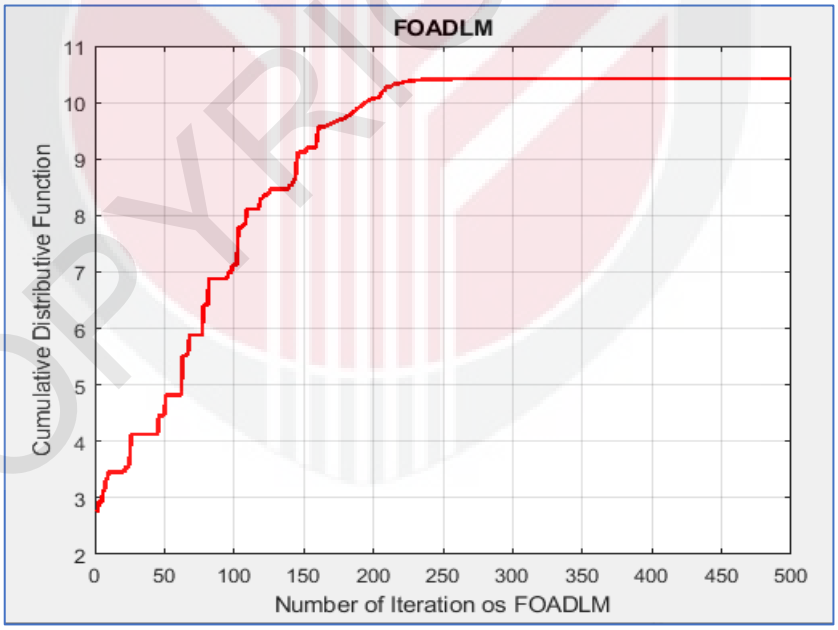


Figure 4.16: Cumulative Distribution Function Analysis of FOADLM

After identifying the distribution of the data in the search space (Figure 4.16), the selected parameters are utilized in the data analysis. The selected parameters update

the network weight and bias value. With the frequent updating of the network parameters, the output deviations are minimized. The discrepancies are calculated by comparing the produced data to the training template. The network consumes a low loss value and achieves good accuracy during training, as shown in Figure 4.17

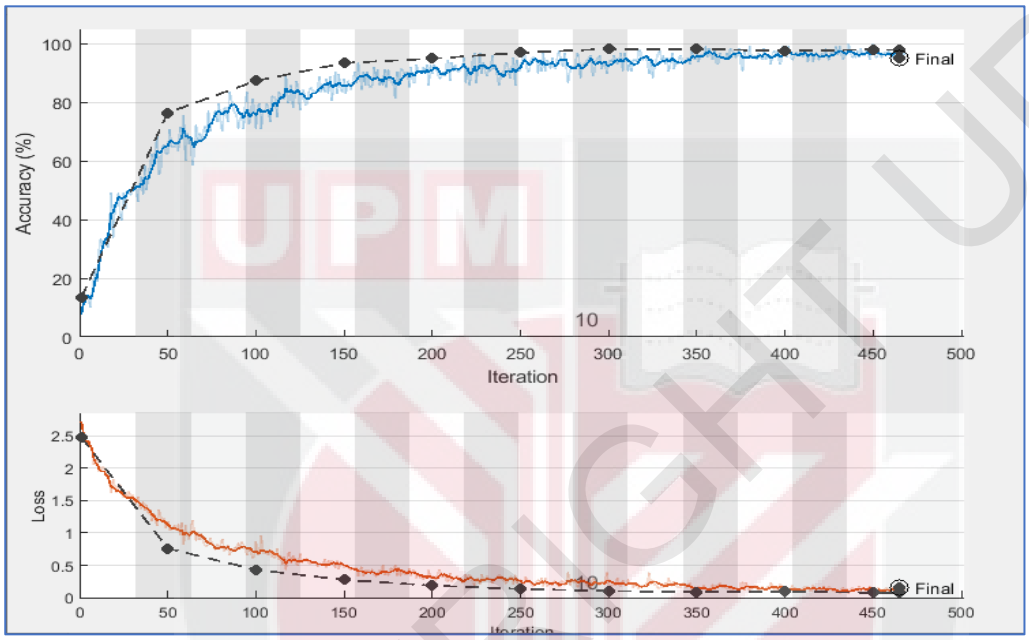


Figure 4.17: Graphical Analysis of the Training Process

Figure 4.17 shows a graphical representation of the training process over time. The output value for the specified input is calculated using 500 iterations in this work. The figure shows that the introduced FOADLM approach has a minimum loss value while iteration increases. The iteration process uses more data that trains frequently, which improves the overall detection accuracy and minimizes the loss value. The utilization of the flock optimization algorithm and working process reduces the accuracy-related issues effectively.

Objective 3: To minimize the computation complexity and maximize the classification accuracy.

This work's third objective is to minimize valuable information loss while analysing the data. During the data analysis, little information is lost while featuring extraction and data analysis. The research issue is overcome by analysing the relationship between each feature while investigating the data. The research uses the Gaussian Filtering approach with the gaussian kernel value that successfully removes the noise from the image. The filtering process eliminates the irrelevant information in the starting stage, reducing the data processing difficulties. More ever, the flock optimization based on selected features and relationship analysis minimizes the information value during the feature extraction process. Then the obtained information loss related to graphical analysis is illustrated in Figure 4.18

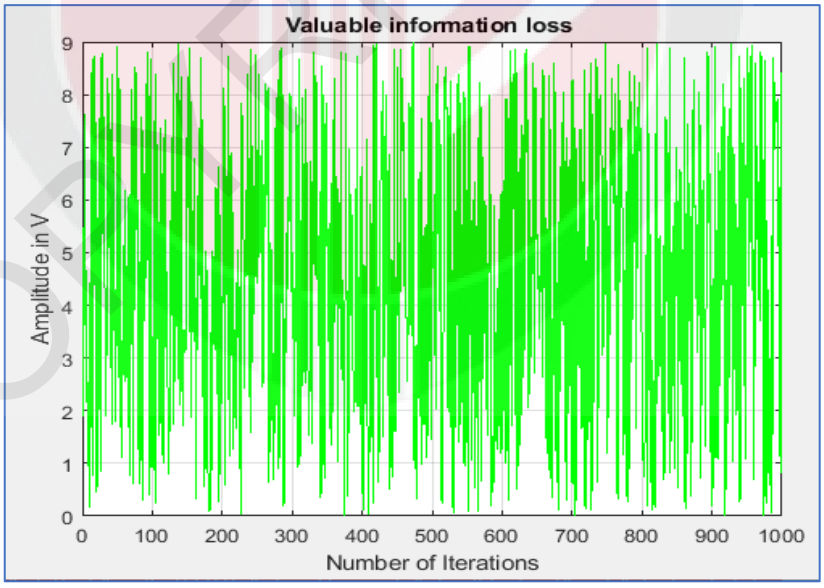


Figure 4.18: Information Loss Analysis

Figure 4.18 illustrates the information loss analysis while classifying people with diabetes. The analysis is performed for 1000 iterations, and the information loss value

is computed for amplitude. The choice of 1000 iterations for training ensures that the model undergoes enough optimization cycles to reach convergence, allowing the weights to stabilize. This number was determined based on trial and error, observing when the loss function plateaued or improved minimally, indicating that further iterations would not significantly enhance performance, thereby serving as the stopping condition. The data distribution clearly shows that the system manages the important data that improves the overall detection efficiency. The data has been further analysed with the help of the Simple Random Sampling (SRS) approach that analyses each data characteristic. According to the characteristics, similar features are clustered and analysed using the Long Short Term Memory network to classify the features with maximum recognition accuracy. The Long Short Term Memory network makes use of a memory cell that records the results of the entire processing, hence speeding up the network's operations. The effective utilization of the flock optimization algorithm predicts the optimized solution with minimum computation cost, shown in Figure 4.19

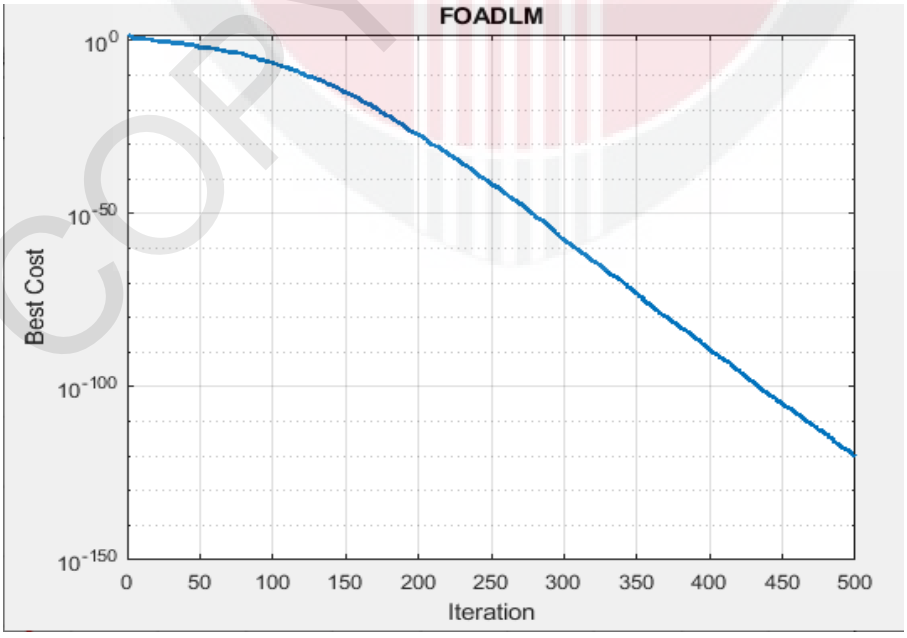


Figure 4.19: Computation Cost Analysis

Figure 4.19 shows that the introduced FOADLM approach attains the minimum computation cost. The network uses the training function that uses the data set to train the features to get the output value. The training process is performed for 500 iterations in which the system ensures a high accuracy value. The frequent training process successfully identifies the output template for every feature. The successive training process reduces the computation difficulties and improves detection efficiency. Therefore, the system consumes minimum computation cost for analysing the output value in the search space. Hence, the Flock Optimization Assimilated Deep Learning Model (FOADLM) successfully resolves data assimilation, information loss, imbalanced data, and accuracy-related issues.

4.9 Summary

Thus, the chapter analyzes the efficiency of the Flock Optimization Assimilated Deep Learning Model (FOADLM) based diabetic detection System. During the analysis, three datasets, such as the PIMA Indian dataset, OhioT1DM Dataset, and Azure diabetic dataset information, are utilized. Data from many sources, including PIMA Indian, is used to evaluate the system's quality. The computed system efficiency is compared with existing methods such as Kezhi Li systems, Deep Neural Network (DNN), Improved Particle swarm optimization optimizing Long short-term memory network (IPSO-LSTM), gravitational search optimized deep learning (GSODL), Tubu optimized sequence modular network (TOSMN), Adaptive Neuro-Fuzzy Inference System (ANFIS), Ensemble Learning Approach (ELA) and DiabDeep (H. Yin et al. 2021). Among the various analyses, the introduced FOADLM approach attains 99.23% accuracy.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Diabetic detection is a vital part of the healthcare system because this helps lower people's risk of developing the disease. The self-assessment process helps continuously monitor the patient's health and minimize the risk factor. According to the self-assessment, several researchers creates the automatic diabetic detection system. The system uses data mining, machine learning, and Artificial Intelligence techniques. These techniques successfully analyze the health data, and abnormal information is predicted. However, the existing systems fail to solve the data assimilation, imbalance data, information loss, and accuracy issues. The research issues are explored with the help of optimized deep-learning approaches. The deep learning concept is utilized according to the Kezhi Li system, and the layer information is changed to improve the overall diabetic detection process. The first layer is worked along with the Flock optimization technique that explores the relationship between the data. This process explores the feature map processed by layer two that reduces the data availability issue. Gaussian Filtering is applied during the analysis to remove the noise that improves the overall data analysis efficiency, and minimizes the imbalance issue. Further, the Long Short Term Memory network is applied to processing the feature map that recognizes diabetic information with minimum computation complexity. Then the system achieves the following objectives

- Resolved data assimilation and imbalanced issues by applying the Gaussian Filtering and Simple Sampling Technique
- Improved the diabetic disease recognition accuracy using the flock-optimized neural model.
- Minimizing the information loss by updating the network parameters using the flock-optimized-based selected parameters.

5.2 Contribution of this study

This study contributes significantly to the development of a versatile and precise method for predicting glucose levels. A significant advancement is the innovative application of data sampling methods such as Flock optimization in the initial layer of a Convolutional Neural Network (CNN) structure. The flock optimization process enhances the data assimilation and reduces the bias by creating the generalization model for analyzing diabetic data. The flock optimization based learning process account the uneven data distribution between the non-diabetic and diabetic data. In addition, the network parameters are updated using flock working process which minimize the deviation between the outputs. The effective utilization of predominate and latent features reduces the classification complexity and ensures the faster convergences. Flock optimization maintains model flexibility by examining correlations between various data attributes and capacity to categorize the data. The sampling-based optimization method in the CNN efficiently captures key characteristics, reducing the difference between actual and forecasted glucose levels. The Flock optimization-based CNN shows improved flexibility and accuracy (99.23%) in glucose prediction as compared to traditional methods.

5.3 Limitation of this study

The suggested methodologies show enhancements in flexibility, accuracy, and efficiency for glucose prediction. However, there are limitations in this study that might be addressed in future research. One restriction is that the evaluation is restricted to the particular dataset employed for training and testing. Further validation on other real-world datasets might enhance the demonstration of the suggested approaches' generalizability. Additionally, both the Flock optimization and Long Short Term Memory optimization methods need tweaking of extra hyperparameters. The study requires comprehensive analysis of the effects of various parameter configurations. Additionally, the computing burden caused by the sampling and optimization methods needs to be assessed, particularly for real-time forecasting. The study does not provide an in-depth review of the model's complexity, training duration, and prediction speed. The optimized CNN-LSTM model's interpretation and explainability should be enhanced. Activation mapping and feature attribution techniques offer insights into the model's process for predicting glucose values. Overcoming these constraints in future research can enhance the effectiveness and relevance of the suggested glucose forecasting system.

The approach also incorporates an Optimized Long Short-Term Memory (LSTM) network to enhance model training efficiency and decrease training duration. This is accomplished by adjusting network parameters based on related information from earlier stages, which helps decrease overfitting to the training data. The optimization procedure also reduces problems associated with data assimilation and imbalance. The Flock optimized CNN and Optimized LSTM network work together to create a versatile and efficient framework for precise and real-time glucose prediction. This

study showcases major scientific contributions in utilizing data sampling strategies and optimization approaches for both CNN and Long Short Term Memory networks.

The model uses the frequent parameter updating process for maintaining the adaptability but leads to various challenges such as required more resources and computational cost to minimize overfitting issues. In addition, the model not generalize to the new set of data. The new data is too sensitive that leads to create the instability while training the data. The instability factor creates the ambiguity during the hyperparameter training, learning rate adjustment and stability management.

5.4 Recommendation for Future work

This study shows promising outcomes; however, several research fields can further develop these findings. Evaluate the Flock-optimized CNN and Long Short Term Memory model on several real-world glucose datasets to enhance its generalizability. Studying the effects of various hyperparameter settings on model performance might offer valuable information on the best setups. This is advisable to analyze the computational complexity and overheads of sampling and optimization strategies, particularly for real-time prediction applications. Future research might aim to enhance model interpretability using techniques such as activation mapping and feature attribution. This can enhance comprehension of the prediction process. Moreover, the refined models might be implemented and assessed on edge devices to show practicality and effectiveness in real-life clinical environments. Integrating contextual variables such as nutrition, activity, and medicines can improve the precision and customization of the glucose forecasting model. Overall, further assessment, refinement, analysis, and validation of the suggested method in future research can enhance its practical use.

REFERENCES

Abbood, S.H., Hamed, H.N.A., Rahim, M.S.M., Rehman, A., Saba, T. & Bahaj, S.A., 2022. Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model. *IEEE Access*, vol. 10, pp. 73079-73086.

Abnoosian, K., Farnoosh, R. and Behzadi, M.H., 2023. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC bioinformatics*, *24*(1), p.337.

Agrawal, S., Tiwari, A. and Goel, I., 2020. Genetically optimized deep neural learning for breast cancer prediction. In *Soft Computing for Problem Solving 2019: Proceedings of SocProS 2019, Volume 2* (pp. 127-139). Springer Singapore.

Aguinis, H. & Solarino, A.M., 2019. Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*, vol. 40, no. 8, pp. 1291-1315.

Ahmed, M.R., Zhang, Y., Feng, Z., Lo, B., Inan, O.T. & Liao, H., 2019. Neuroimaging and Machine Learning for Dementia Diagnosis: Recent Advancements and Future Prospects. *IEEE Reviews in Biomedical Engineering*, 12, pp.19-33. DOI: 10.1109/RBME.2018.2886237.

Ahmed, U., Issa, G.F., Khan, M.A., Aftab, S., Khan, M.F., Said, R.A. & Ahmad, M., 2022. Prediction of diabetes empowered with fused machine learning. *IEEE Access*, vol. 10, pp. 8529-8538.

Ahmedt-Aristizabal, D., Rehman, N., Tafreshi, R., Choupan, J., Tran, Y. and Breakspear, M., 2021. Identification of children at risk of schizophrenia via deep learning and EEG responses. *IEEE Journal of Biomedical and Health Informatics*, 25(1), pp.69-76. doi: 10.1109/JBHI.2020.2984238.

Akhtar, M.A.K., Pradhan, D.K., Kumar Chatterjee, R., Chakraborty, F., Kumar, M., Verma, S., ... & García Arenas, M.I., 2024. FNN for diabetic prediction using oppositional whale optimization algorithm. in *IEEE Access*, vol. 12, pp. 20396-20408, 2024, doi: 10.1109/ACCESS.2024.3357993

Akshatha Rao, M., Lamani, D., Bhandarkar, R. & Manjunath, T.C., 2014. Automated detection of diabetic retinopathy through image feature extraction. *International Conference on Advances in Electronics Computers and Communications*.

Alhussan, A.A., Abdelhamid, A.A., Towfek, S.K., Ibrahim, A., Eid, M.M., Khafaga, D.S. & Saraya, M.S., 2023. Classification of diabetes using feature selection and hybrid Al-Biruni earth radius and dipper throated optimization. *Diagnostics*, 13(12), p. 2038.

Ali, F., El-Sappagh, S., Islam, S.M.R., Kwak, D., Ali, A., Imran, M. & Kwak, K.S., 2020. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, vol. 63, pp. 208-222.

Ali, H., Umander, J., Rohlén, R. & Grönlund, C., 2020. A Deep Learning Pipeline for Identification of Motor Units in Musculoskeletal Ultrasound. *IEEE Access*, 8, pp.170595-170608. DOI: 10.1109/ACCESS.2020.3023495.

Aliyu, H.A., Muritala, I.O., Bello-Salau, H., Mohammed, S., Onumanyi, A.J. and Ajayi, O.O., 2024. Optimizing machine learning algorithms for diabetes data: A metaheuristic approach to balancing and tuning classifiers parameters. *Franklin Open*, *8*, p.100153.

Al-Tawil, M., Mahafzah, B.A., Al Tawil, A. & Aljarah, I., 2023. Bio-Inspired Machine Learning Approach to Type 2 Diabetes Detection. *Symmetry*, vol. 15, no. 3, p. 764.

AlZubi, A.A., Alarifi, A. & Al-Maitah, M., 2020. Deep brain simulation wearable IoT sensor device based Parkinson brain disorder detection using heuristic tubu optimized sequence modular neural network. *Measurement*, vol. 161, p. 107887.

Amarbayasgalan, T., Park, K.H., Lee, J.Y. & Ryu, K.H., 2019. Reconstruction error-based deep neural networks for coronary heart disease risk prediction. *PLoS One*, vol. 14, no. 12, p. e0225991.

An, X., Zhang, Y., Cao, Y., Chen, J., Qin, H. & Yang, L., 2020. Punicalagin protects diabetic nephropathy by inhibiting pyroptosis based on TXNIP/NLRP3 pathway. *Nutrients*, vol. 12, no. 5, p. 1516.

Arokiasamy, P., Salvi, S. & Selvamani, Y., 2021. Global burden of diabetes mellitus: prevalence, pattern, and trends. *Handbook of Global Health*, pp. 495-538.

Ashour, A.F., Fouda, M.M., Fadlullah, Z.M. & Ibrahem, M.I., 2024. Optimized Neural Networks for Diabetes Classification Using Pima Indians Diabetes Database. In: *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, April, pp. 1–7. IEEE.

Atteia, G., Abdel Samee, N., El-Kenawy, E.S.M. & Ibrahim, A., 2022. CNN-hyperparameter optimization for diabetic maculopathy diagnosis in optical coherence tomography and fundus retinography. *Mathematics*, 10(18), p. 3274.

Awuchi, C.G., Echeta, C.K. & Igwe, V.S., 2020. Diabetes and the nutrition and diets for its prevention and treatment: a systematic review and dietetic perspective. *Health Sciences Research*, vol. 6, no. 1, pp. 5-19.

Aynalem, S.B. & Zeleke, A.J., 2018. Prevalence of Diabetes Mellitus and Its Risk Factors among Individuals Aged 15 Years and Above in Mizan-Aman Town, Southwest Ethiopia, 2016: A Cross Sectional Study. *International Journal of Endocrinology*, 2018, Article ID 9317987, 7 pages. DOI: https://doi.org/10.1155/2018/9317987.

Bahaidarah, M., Rekabi-Bana, F., Marjanovic, O., & Arvin, F. (2024). Swarm flocking using optimisation for a self-organised collective motion. *Swarm and Evolutionary Computation*, 86, 101491.

Bak, J.C., Mul, D., Serné, E.H., de Valk, H.W., Sas, T.C., Geelhoed-Duijvestijn, P.H., & Verheugt, C.L., 2021. DPARD: rationale, design and initial results from the Dutch national diabetes registry. *BMC Endocrine Disorders*, vol. 21, no. 1, pp. 1-10.

Balwan, W.K. & Kour, S., 2021. Lifestyle Diseases: The Link between Modern Lifestyle and threat to public health. *Saudi J Med Pharm Sci*, vol. 7, no. 4, pp. 179-184.

Bansal, J. C., Sethi, N., Anicho, O., & Nagar, A. (2023). Drone flocking optimization using NSGA-II and principal component analysis. *Swarm Intelligence*, *17*(1), 63-87.

Bansode, B. & Prasad, J.B., 2022. Burden of comorbidities among diabetic patients in Latur, India. *Clinical Epidemiology and Global Health*, vol. 13, p. 100957.

Barber, R.M., Sorensen, R.J., Pigott, D.M., Bisignano, C., Carter, A., Amlag, J.O., Collins, J.K., et al., 2022. Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: a statistical analysis. *The Lancet*, vol. 399, no. 10344, pp. 2351-2380.

Barke, A., Korwisi, B., Jakob, R., Konstanjsek, N., Rief, W. & Treede, R.-D., 2022. Classification of chronic pain for the International Classification of Diseases (ICD-11): results of the 2017 international World Health Organization field testing. *Pain*, vol. 163, no. 2, p. e310.

Barreto, F.C., Barreto, D.V., Massy, Z.A. & Drüeke, T.B., 2019. Strategies for phosphate control in patients with CKD. *Kidney International Reports*, vol. 4, no. 8, pp. 1043-1056.

Baruah, H.S., Thakur, J., Sarmah, S. & Hoque, N., 2020. A Feature Selection Method using PSO-MI. *2020 International Conference on Computational Performance Evaluation (ComPE)*, Shillong, India, pp.280-284. DOI: 10.1109/ComPE49325.2020.9200034.

Basu, S., 2020. Non-communicable disease management in vulnerable patients during Covid-19. *Indian J Med Ethics*, vol. 2, pp. 103-105.

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. Frontiers in psychology, 12, 675558.

Bhaskar, N., Bairagi, V., Boonchieng, E. and Munot, M.V., 2023. Automated detection of diabetes from exhaled human breath using deep hybrid architecture. *IEEE Access*, *11*, pp.51712-51722.

Bilal, A., Zhu, L., Deng, A., Lu, H. & Wu, N., 2022. AI-based automatic detection and classification of diabetic retinopathy using U-Net and deep learning. *Symmetry*, vol. 14, no. 7, p. 1427.

Bishop, A.N. & Del Moral, P., 2023. On the mathematical theory of ensemble (linear-Gaussian) Kalman–Bucy filtering. *Mathematics of Control, Signals, and Systems*, 35(4), pp.835-903.

Bode, N.W., Franks, D.W. and Wood, A.J., 2011. Limited interactions in flocks: relating model simulations to empirical data. *Journal of The Royal Society Interface*, *8*(55), pp.301-304.

Boucsein, A., Kamstra, K. & Tups, A., 2021. Central signalling cross-talk between insulin and leptin in glucose and energy homeostasis. *Journal of Neuroendocrinology*, vol. 33, no. 4, p. e12944.

Brownlee, J., 2019. How to choose a feature selection method for machine learning. *Machine Learning Mastery*, *10*, pp.1-7.

Brož, J., Malinovská, J., Nunes, M.A., Kučera, K., Rožeková, K., Žejglicová, K., Urbanová, J., Jenšovský, M., Brabec, M. & Lustigová, M., 2020. Prevalence of diabetes and prediabetes and its risk factors in adults aged 25–64 in the Czech Republic: A cross-sectional study. *Diabetes Research and Clinical Practice*, vol. 170, p. 108470.

Byra, M., Dobruch-Sobczak, K., Klimonda, Z., Piotrzkowska-Wroblewska, H. & Litniewski, J., 2021. Early Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer Sonography Using Siamese Convolutional Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, 25(3), pp.797-805. DOI: 10.1109/JBHI.2020.3008040.

Cenitta, D., Arjunan, R.V. and Prema, K.V., 2022. Ischemic heart disease prediction using optimized squirrel search feature selection algorithm. *IEEE Access*, 10, pp.122995-123006.

Chatterjee, R., Akhtar, M.A.K., Pradhan, D.K., Chakraborty, F., Kumar, M., Verma, S., ... & García-Arenas, M., 2024. FNN for diabetic prediction using oppositional whale optimization algorithm. *IEEE Access,* vol. 12, pp. 20396-20408, 2024, doi: 10.1109/ACCESS.2024.3357993.

Chen, Y., Gao, Y., Li, K., Zhao, L. & Zhao, J., 2020. Vertebrae Identification and Localization Utilizing Fully Convolutional Networks and a Hidden Markov Model. *IEEE Transactions on Medical Imaging*, 39(2), pp.387-399. DOI: 10.1109/TMI.2019.2927289.

Cherian, R.P., Thomas, N. and Venkitachalam, S., 2020. Weight optimized neural network for heart disease prediction using hybrid lion plus particle swarm algorithm. *Journal of Biomedical Informatics*, 110, p.103543.

Chong, J., Tjurin, P., Niemelä, M., Jämsä, T. and Farrahi, V., 2021. Machine-learning models for activity class prediction: a comparative study of feature selection and classification algorithms. *Gait & Posture*, 89, pp.45-53.

Chou, F. I., Huang, T. H., Yang, P. Y., Lin, C. H., Lin, T. C., Ho, W. H., & Chou, J. H. (2021). Controllability of Fractional-Order Particle Swarm Optimizer and Its Application in the Classification of Heart Disease. *Applied Sciences*, *11*(23), 11517.

Dagher, K.E. & Haggege, J., 2024. Enhancement of the blood glucose level for diabetic patients based on an adaptive auto-tuned PID controller via meta-heuristic methods. *International Journal of Intelligent Engineering & Systems*, 17(3).

Daliri, M.R., 2012. Feature selection using binary particle swarm optimization and support vector machines for medical diagnosis. *Biomedizinische Technik/Biomedical Engineering*, 57(5), pp.395-402.

David, D. and Namboodiri, S., 2020. Improvement of framework for the grouping of CA diseases by investigating big data. *Artech Journal of Engineering Applied Technology*, 1, pp.7-14.

de La Torre, J., Valls, A. and Puig, D., 2020. A deep learning interpretable classifier for diabetic retinopathy disease grading. *Neurocomputing*, 396, pp.465-476.

Dejkhamron, P., Santiprabhob, J., Likitmaskul, S., Deerochanawong, C., Rawdaree, P., Tharavanij, T. et al., 2022. Young-onset diabetes patients in Thailand: data from Thai Type 1 Diabetes and Diabetes diagnosed Age before 30 years Registry, Care and Network (T1DDAR CN). *Journal of Diabetes Investigation*, 13(5), pp.796-809.

Du, C., Du, C., Huang, L. and He, H., 2019. Reconstructing perceived images from human brain activities with Bayesian deep multiview learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8), pp.2310-2323. doi: 10.1109/TNNLS.2018.2882456.

Du, Y., Dennis, B., Rhodes, S.L., Sia, M., Ko, J., Jiwani, R. and Wang, J., 2020. Technology-assisted self-monitoring of lifestyle behaviours and health indicators in diabetes: qualitative study. *JMIR Diabetes*, 5(3), p.e21183.

Dutta, A., Hasan, M.K., Ahmad, M., Awal, M.A., Islam, M.A., Masud, M. and Meshref, H., 2022. Early prediction of diabetes using an ensemble of machine learning models. *International Journal of Environmental Research and Public Health*, 19(19), p.12378.

Ebrahimi-Ghahnavieh, A., Luo, S. & Chiong, R., 2019. Transfer Learning for Alzheimer's Disease Detection on MRI Images. In *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, Bali, Indonesia, 2019, pp. 133-138. doi: 10.1109/ICIAICT.2019.8784845.

Ezzat, D., Hassanien, A.E. and Ella, H.A., 2021. An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. *Applied Soft Computing*, *98*, p.106742.

Facondo, P., Di Lodovico, E., Delbarba, A., Anelli, V., Pezzaioli, L.C., Filippini, E. & Ferlin, A., 2022. The impact of diabetes mellitus type 1 on male fertility: Systematic review and meta-analysis. *Andrology*, 10(3), pp.426-440.

Fatima, N., Liu, L., Hong, S. & Ahmed, H., 2020. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access*, 8, pp.150360-150376. DOI: 10.1109/ACCESS.2020.3016715.

Febrian, M.E., Ferdinan, F.X., Sendani, G.P., Suryanigrum, K.M. & Yunanda, R., 2023. Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216, pp.21-30.

Fitriyani, N.L., Syafrudin, M., Alfian, G. & Rhee, J., 2019. Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. *IEEE Access*, 7, pp.144777-144789. DOI: 10.1109/ACCESS.2019.2945129.

Fouad, H., Hassanein, A.S., Soliman, A.M. & Al-Feel, H., 2020. Analyzing patient health information based on IoT sensor with AI for improving patient assistance in the future direction. *Measurement*, 159, p.107757.

Ganie, S.M., Pramanik, P.K.D., Bashir Malik, M., Mallik, S. & Qin, H., 2023. An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics*, 14, p. 1252159.

Garg, S.K., Rodbard, D., Hirsch, I.B. & Forlenza, G.P., 2020. Managing new-onset type 1 diabetes during the COVID-19 pandemic: challenges and opportunities. *Diabetes Technology & Therapeutics*, 22(6), pp.431-439.

Gedam, A. N., Ajalkar, D. A., & Rumale, A. S. (2024). Lung nodule detection using Eyrie Flock-based Deep Convolutional Neural Network. *Intelligent Decision Technologies*, *18*(3), 1651-1673.

Geem, D. et al., 2024. Progression of Pediatric Crohn's Disease Is Associated with Anti–Tumor Necrosis Factor Timing and Body Mass Index Z-Score Normalization. *Clinical Gastroenterology and Hepatology*, 22(2), pp.368-376.

Geraghty, T. & Laporta, G., 2019. Current health and economic burden of chronic diabetic osteomyelitis. *Expert Review of Pharmacoeconomics & Outcomes Research*, 19(3), pp.279-286.

Ghabousian, R., Farhang, Y., Majidzadeh, K. & Babazadeh Sangar, A., 2024. Hybrid of particle swarm optimization algorithm and fuzzy system for diabetes diagnosis. *International Journal of Nonlinear Analysis and Applications*, 15(2), pp. 39–46.

Gregory, J.M., Slaughter, J.C., Duffus, S.H., Smith, T.J., LeStourgeon, L.M., Jaser, S.S., McCoy, A.B., et al., 2021. COVID-19 severity is tripled in the diabetes community: a prospective analysis of the pandemic's impact in type 1 and type 2 diabetes. *Diabetes Care*, 44(2), pp.526-532.

Gupta, D., Sundaram, S., Khanna, A., Hassanien, A.E. & De Albuquerque, V.H.C., 2018. Improved diagnosis of Parkinson's disease using optimized crow search algorithm. *Computers & Electrical Engineering*, 68, pp.412-424.

Hajam, Y.A., Rani, R., Malik, J.A., Pandita, A., Sharma, R. & Kumar, R., 2023. Diabetes Mellitus: Signs and Symptoms, Epidemiology, Current Prevention, Management Therapies, and Treatments. In: Antidiabetic Potential of Plants in the Era of Omics. Apple Academic Press, pp.31-77.

Haque, F., Reaz, M.B.I., Chowdhury, M.E.H., Hashim, F.H., Arsad, N. & Ali, S.H.M., 2021. Diabetic Sensorimotor Polyneuropathy Severity Classification Using Adaptive Neuro Fuzzy Inference System. *IEEE Access*, 9, pp.7618-7631. DOI: 10.1109/ACCESS.2020.3048742.

Hassan, G.S., Ali, N.J., Abdulsahib, A.K., Mohammed, F.J. & Gheni, H.M., 2023. A missing data imputation method based on salp swarm algorithm for diabetes disease. *Bulletin of Electrical Engineering and Informatics*, 12(3), pp. 1700–1710.

He, J. & Wang, Y., 2020. Blood glucose concentration prediction based on kernel canonical correlation analysis with particle swarm optimization and error compensation. *Computer Methods and Programs in Biomedicine*, 196, p.105574.

Hegde, S. & Mundada, M.R., 2021. Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach. *International Journal of Pervasive Computing and Communications*, 17(1), pp.20-36.

Hereford, J. and Blum, C., 2011, October. FlockOpt: A new swarm optimization algorithm based on collective behaviour of starling birds. In *2011 Third World Congress on Nature and Biologically Inspired Computing* (pp. 17-22). IEEE.

Hina, A. & Saadeh, W., 2022. Noninvasive blood glucose monitoring systems using near-infrared technology—A review. *Sensors*, 22(13), p.4855.

Homayouni, A., Liu, T. & Thieu, T., 2022. Diabetic retinopathy prediction using Progressive Ablation Feature Selection: A comprehensive classifier evaluation. *Smart Health*, 26, p.100343.

Hou, L., Zhu, J., Kwok, J., Gao, F., Qin, T. & Liu, T.Y., 2019. Normalization helps training of quantized lstm. *Advances in Neural Information Processing Systems*, 32.

Howlader, K.C., Satu, M.S., Awal, M.A., Islam, M.R., Islam, S.M.S., Quinn, J.M. & Moni, M.A., 2022. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. *Health Information Science and Systems*, 10(1), p.2.

Huang, M.L. & Chou, Y.C., 2019. Combining a gravitational search algorithm, particle swarm optimization, and fuzzy rules to improve the classification performance of a feed-forward neural network. *Computer Methods and Programs in Biomedicine*, 180, p.105016.

Iacopi, E., Pieruzzi, L., Riitano, N., Abbruzzese, L., Goretti, C. & Piaggesi, A., 2023. The weakness of the strong sex: Differences between men and women affected by diabetic foot disease. *The International Journal of Lower Extremity Wounds*, 22(1), pp.19-26.

Ibuki, T., Wilson, S., Yamauchi, J., Fujita, M. & Egerstedt, M., 2020. Optimization-Based Distributed Flocking Control for Multiple Rigid Bodies. *IEEE Robotics and Automation Letters*, 5(2), pp.1891-1898. DOI: 10.1109/LRA.2020.2969950.

Islam, M.M., Ferdousi, R., Rahman, S. & Bushra, H.Y., 2020. Likelihood prediction of diabetes at early stage using data mining techniques. In: *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer, Singapore, pp.113-125.

Izci, D., Ekinci, S. & Mirjalili, S., 2023. Optimal PID plus second-order derivative controller design for AVR system using a modified Runge Kutta optimizer and Bode's ideal reference model. *International Journal of Dynamics and Control*, 11(3), pp.1247-1264.

Javeria, A., Sharif, M. & Yasmin, M., 2016. A Review on Recent Developments for Detection of Diabetic Retinopathy. *Scientifica*, 2016, Article ID 6838976, 20 pages. DOI: https://doi.org/10.1155/2016/6838976.

Kakitapalli, Y., Ampolu, J., Madasu, S.D. & Kumar, M.L.S., 2020. Detailed review of chronic kidney disease. *Kidney Diseases*, 6(2), pp.85-91.

Kamalraj, R., Neelakandan, S., Kumar, M.R., Rao, V.C.S., Anand, R. & Singh, H., 2021. Interpretable filter based convolutional neural network (IF-CNN) for glucose prediction and classification using PD-SS algorithm. *Measurement*, 183, p.109804.

Kanyongo, W. & Ezugwu, A.E., 2023. Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. *Informatics in Medicine Unlocked*, p.101232.

Karale, A., Lazarova, M., Koleva, P. & Poulkov, V., 2020. A Hybrid PSO-MiLOF Approach for Outlier Detection in Streaming Data. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, Milan, Italy, 2020, pp. 474-479. doi: 10.1109/TSP49548.2020.9163430.

Karayılan, T. & Kılıç, Ö., 2017. Prediction of heart disease using neural network. *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkey, pp.719-723. DOI: 10.1109/UBMK.2017.8093512.

Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L. and Salcudean, S.E., 2020. Deep learning-based Gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. *IEEE Journal of Biomedical and Health Informatics*, 24(5), pp.1413-1426. doi: 10.1109/JBHI.2019.2944643.

Kaur, S. et al., 2020. Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. *IEEE Access*, 8, pp.228049-228069. DOI: 10.1109/ACCESS.2020.3042273.

Kaur, S., Aggarwal, H. & Rani, R., 2020. Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease. *Machine Vision and Applications*, 31(32). DOI: https://doi.org/10.1007/s00138-020-01078-1.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaver, N., Vlahavasa, I. & Chouvarda, I., 2017. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, pp.104-116.

Khan, H.N., Shahid, A.R., Raza, B., Dar, A.H. & Alquhayz, H., 2019. Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network. *IEEE Access*, 7, pp.165724-165733. DOI: 10.1109/ACCESS.2019.2953318.

Khan, R.M.M., Chua, Z.J.Y., Tan, J.C., Yang, Y., Liao, Z. & Zhao, Y., 2019. From pre-diabetes to diabetes: diagnosis, treatments and translational research. *Medicina*, 55(9), p.546.

Kholwal, R., 2023. Enhancing Early-Stage Diabetes Prediction Using Data Mining Algorithms and Normalisation Techniques. *International Journal of Advances in Engineering & Technology*, 16(5), pp.302-313.

Kumar, A., Kaur, A., Singh, P., Driss, M. & Boulila, W., 2023. Efficient Multiclass Classification Using Feature Selection in High-Dimensional Datasets. *Electronics*, 12(10), p.2290.

Kumar, V., Recupero, D.R., Riboni, D. & Helaoui, R., 2021. Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes. *IEEE Access*, 9, pp.7107-7126. DOI: 10.1109/ACCESS.2020.3043221.

Kumari, S., Laxmikant, S.D., Sonika, B. & Khanal, S., 2022. Efficacy of Integrated Ayurveda treatment protocol in type 2 Diabetes Mellitus–A case report. *Journal of Ayurveda and Integrative Medicine*, 13(1), p.100512.

Lalama, Z., Boulfekhar, S. & Semechedine, F., 2022. Localization optimization in wsns using meta-heuristics optimization algorithms: A survey. *Wireless Personal Communications*, pp.1-24.

Langerman, C., Forbes, A. & Robert, G., 2022. The experiences of insulin use among older people with Type 2 diabetes mellitus: A thematic synthesis. *Primary Care Diabetes*.

Laudani, A., Fulginei, F.R., Lozito, G.M. & Salvini, A., 2014. Swarm/flock optimization algorithms as continuous dynamic systems. *Applied Mathematics and Computation*, 243, pp.670-683.

Le, T.M., Vo, T.M., Pham, T.N. & Dao, S.V.T., 2020. A novel wrapper-based feature selection for early diabetes prediction enhanced with a metaheuristic. *IEEE Access*, 9, pp. 7869–7884.

Li, K., Daniels, J., Liu, C., Herrero, P. & Georgiou, P., 2019. Convolutional recurrent neural networks for glucose prediction. *IEEE Journal of Biomedical and Health Informatics*, 24(2), pp.603-613.

Li, K., Liu, C., Zhu, T., Herrero, P. & Georgiou, P., 2019. GluNet: A Deep Learning Framework for Accurate Glucose Forecasting. *IEEE Journal of Biomedical and Health Informatics*, PP(c), pp.1.

Li, X., Zhang, J. & Safara, F., 2023. Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm. *Neural Processing Letters*, 55(1), pp. 153–169.

Li, Y., Li, H. & Yao, H., 2018. Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017. *Computational and Mathematical Methods in Medicine*, 2018, Article ID 7207151, 8 pages. DOI: https://doi.org/10.1155/2018/7207151.

Li, Y.H., Yeh, N.N., Chen, S.J. & Chung, Y.C., 2019. Computer-Assisted Diagnosis for Diabetic Retinopathy Based on Fundus Images Using Deep Convolutional Neural Network. *Mobile Information Systems*, 2019, Article ID 6142839, 14 pages. DOI: https://doi.org/10.1155/2019/6142839.

Liang, Q. et al., 2019. Weakly Supervised Biomedical Image Segmentation by Reiterative Learning. *IEEE Journal of Biomedical and Health Informatics*, 23(3), pp.1205-1214. DOI: 10.1109/JBHI.2018.2850040.

Liang, Y.Q., Zhou, R., Chen, H.W., Cao, B.F., Fan, W.D., Liu, K., ... & Zou, M.C., 2023. Associations of blood biomarkers with arterial stiffness in patients with diabetes mellitus: A population-based study. *Journal of Diabetes*.

Liccardo, D., Cannavo, A., Spagnuolo, G., Ferrara, N., Cittadini, A., Rengo, C. & Rengo, G., 2019. Periodontal disease: a risk factor for diabetes and cardiovascular disease. *International Journal of Molecular Sciences*, 20(6), p.1414.

Linn, W., Persson, M., Rathsman, B., Ludvigsson, J., Lind, M., Andersson Franko, M. & Nyström, T., 2023. Estimated glucose disposal rate is associated with retinopathy and kidney disease in young people with type 1 diabetes: a nationwide observational study. *Cardiovascular Diabetology*, 22(1), pp.1-12.

Lozito, G.M. & Salvini, A., 2019. Swarm intelligence based approach for efficient training of regressive neural networks. *Neural Computing and Applications*, pp.1-12.

Lu, C., Koyuncu, C., Corredor, G., Prasanna, P., Leo, P., Wang, X., Janowczyk, A. et al., 2021. Feature-driven local cell graph (FLocK): New computational pathology-based descriptors for prognosis of lung cancer and HPV status of oropharyngeal cancers. *Medical Image Analysis*, 68, p.101903.

Ma, L., Ma, C., Liu, Y. & Wang, X., 2019. Thyroid Diagnosis from SPECT Images Using Convolutional Neural Network with Optimization. *Computational Intelligence and Neuroscience*, 2019, Article ID 6212759, 11 pages. DOI: https://doi.org/10.1155/2019/6212759.

Maddumala, V.R., 2020. Big Data-Driven Feature Extraction and Clustering Based on Statistical Methods. *Traitement du Signal*, 37(3).

Mandal, N., Grambergs, R., Mondal, K., Basu, S.K., Tahia, F. & Dagogo-Jack, S., 2021. Role of ceramides in the pathogenesis of diabetes mellitus and its complications. *Journal of Diabetes and its Complications*, 35(2), p.107734.

Maniruzzaman, Md, Rahman, Md J., Ahammed, B. & Abedin, Md M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems*, 8, pp.1-14.

Maule, P., Shete, A., Wani, K. & Dawange, A., 2015. A Review on GLCM feature extraction in Retinal Image. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 4(9).

Menaouer, B., Dermane, Z., El Houda Kebir, N. & Matta, N., 2022. Diabetic Retinopathy Classification Using Hybrid Deep Learning Approach. *SN Computer Science*, 3(5), p.357.

Mengarelli, A., Tigrini, A., Verdini, F., Rabini, R.A. & Fioretti, S., 2023. Multiscale Fuzzy Entropy Analysis of Balance: Evidences of Scale-Dependent Dynamics on Diabetic Patients with and without Neuropathy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, pp.1462-1471.

Mishra, S., Chaudhury, P., Mishra, B. & Tripathy, H., 2016. An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction. *Proceedings of the ACM*, pp.1-3. DOI: 10.1145/2905055.2905100.

Mishra, S., Tripathy, H.K., Mallick, P.K., Bhoi, A.K. & Barsocchi, P., 2020. EAGA-MLP—an enhanced and adaptive hybrid classification model for diabetes diagnosis. *Sensors*, 20(14), p. 4036.

Mohan, S., Thirumalai, C. & Srivastava, G., 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, 7, pp.81542-81554. DOI: 10.1109/ACCESS.2019.2923707.

Mortazavi, R., Mortazavi, S. & Troncoso, A., 2021. Wrapper-based feature selection using regression trees to predict intrinsic viscosity of polymer. *Engineering with Computers*, pp.1-13.

Mousa, A., Mustafa, W., Marqas, R.B. & Mohammed, S.H., 2023. A comparative study of diabetes detection using the Pima Indian diabetes database. *Journal of Duhok University*, 26(2), pp. 277–288.

MuthuKumar, M.S. & Jayakumar, M., 2024. A novel deep learning integrated optimization-based diabetes detection (DeO-DiDet) system using IoT. *Educational Administration: Theory and Practice*, 30(5), pp. 3728–3740.

Nadesh, R.K. & Arivuselvan, K., 2020. Type 2: diabetes mellitus prediction using deep neural networks classifier. *International Journal of Cognitive Computing in Engineering*, 1, pp.55-61.

Nahas, H., Au, J.S., Ishii, T., Yiu, B.Y.S., Chee, A.J.Y. & Yu, A.C.H., 2020. A Deep Learning Approach to Resolve Aliasing Artifacts in Ultrasound Color Flow Imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(12), pp.2615-2628. DOI: 10.1109/TUFFC.2020.3001523.

Nandan, D., Kanungo, J. & Mahajan, A., 2024. An error-efficient Gaussian filter for image processing by using the expanded operand decomposition logarithm multiplication. *Journal of Ambient Intelligence and Humanized Computing*, pp.1-8.

Naveena, S. & Bharathi, A., 2022. A new design of diabetes detection and glucose level prediction using moth flame-based crow search deep learning. *Biomedical Signal Processing and Control*, 77, p.103748.

Naz, H. & Ahuja, S., 2020. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19, pp. 391–403.

Olisah, C.C., Smith, L. & Smith, M., 2022. Diabetes mellitus prediction and diagnosis from a data pre-processing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, p.106773.

Ornoy, A., Becker, M., Weinstein-Fudim, L. & Ergaz, Z., 2021. Diabetes during pregnancy: A maternal disease complicating the course of pregnancy with long-term deleterious effects on the offspring. A clinical review. *International Journal of Molecular Sciences*, 22(6), p.2965.

Özbay, E., 2023. An active deep learning method for diabetic retinopathy detection in segmented fundus images using artificial bee colony algorithm. *Artificial Intelligence Review*, 56(4), pp.3291-3318.

Padhi, S., Nayak, A.K. & Behera, A., 2020. Type II diabetes mellitus: a review on recent drug based therapeutics. *Biomedicine & Pharmacotherapy*, 131, p.110708.

Park, S., Lee, S.J., Weiss, E. & Motai, Y., 2016. Intra- and Inter-Fractional Variation Prediction of Lung Tumors Using Fuzzy Deep Learning. *IEEE Journal of Translational Engineering in Health and Medicine*, 4, pp.1-12. DOI: 10.1109/JTEHM.2016.2516005.

Pham, T.D., Wårdell, K., Eklund, A. & Salerud, G., 2019. Classification of short time series in early Parkinson's disease with deep learning of fuzzy recurrence plots. *IEEE/CAA Journal of Automatica Sinica*, 6(6), pp.1306-1317. DOI: 10.1109/JAS.2019.1911774.

Pluta, A., Marzec, A. & Kobus, E., 2021. Some aspects of preparing diabetic patients for self-care.

Polat, K. & Güneş, S., 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, 17(4), pp.702-710.

Popping, S., Bade, D., Boucher, C., van der Valk, M., El-Sayed, M., Sigurour, O., Sypsa, V., Morgan, T., Gamkrelidze, A., Mukabatsinda, C. & Deuffic-Burban, S., 2019. The global campaign to eliminate HBV and HCV infection: International Viral Hepatitis Elimination Meeting and core indicators for development towards the 2030 elimination goals. *Journal of Virus Eradication*, 5(1), pp.60-66.

Pradhan, G., Thapa, G., Pradhan, R., Khandelwal, B., Panigrahi, R., Bhoi, A.K. & Barsocchi, P., 2024. Optimized forest framework with a binary multineighborhood artificial bee colony for enhanced diabetes mellitus detection. *International Journal of Computational Intelligence Systems*, 17(1), p. 194.

Preiya, V.S. & Kumar, V.A., 2023. Deep Learning-Based Classification and Feature Extraction for Predicting Pathogenesis of Foot Ulcers in Patients with Diabetes. *Diagnostics*, 13(12), p.1983.

Qi, C., Mao, X., Zhang, Z. and Wu, H., 2017. Classification and differential diagnosis of diabetic nephropathy. *Journal of Diabetes Research*, 2017, Article ID 8637138. Available at: https://doi.org/10.1155/2017/8637138.

Rahim, S.S., Palade, V. & Shuttleworth, J., 2016. Automatic screening and classification of diabetic retinopathy and maculopathy using fuzzy image processing. *Brain Inf*, 3, p.249. DOI: https://doi.org/10.1007/s40708-016-0045-3.

Rahman, K.K.M., Nasor, M. & Imran, A., 2022. Automatic Screening of Diabetic Retinopathy Using Fundus Images and Machine Learning Algorithms. *Diagnostics*, 12(9), p.2262.

Ramesh, S. & Vydeki, D., 2020. Recognition and classification of paddy leaf diseases using Optimized Deep Neural network with Jaya algorithm. *Information Processing in Agriculture*, 7(2), pp.249-260.

Rasta, S.H., Partovi, M.E., Seyedarabi, H. & Javadzadeh, A., 2015. A Comparative Study on Preprocessing Techniques in Diabetic Retinopathy Retinal Images: Illumination Correction and Contrast Enhancement. *J Med Signals Sens*, 5(1), pp.40-48.

Ravishankar, S., Jain, A. & Mittal, A., 2009. Automated feature extraction for early detection of diabetic retinopathy in fundus images. *IEEE Conference on Computer Vision and Pattern Recognition*.

Rehman, A., Athar, A., Khan, M.A., Abbas, S., Fatima, A. & Saeed, A., 2020. Modelling, simulation, and optimization of diabetes type II prediction using deep extreme learning machine. *Journal of Ambient Intelligence and Smart Environments*, 12(2), pp. 125–138.

Rehman, A.U., Islam, A. & Belhaouari, S.B., 2020. Multi-Cluster Jumping Particle Swarm Optimization for Fast Convergence. *IEEE Access*, vol. 8, pp. 189382-189394. doi: 10.1109/ACCESS.2020.3031003.

Robertson, G., Lehmann, E.D., Sandham, W. & Hamilton, D., 2011. Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. *Journal of Electrical and Computer Engineering*, 2011, Article ID 681786, 11 pages. DOI: https://doi.org/10.1155/2011/681786.

Rufo, D.D., Debelee, T.G., Ibenthal, A. & Negera, W.G., 2021. Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics*, 11(9), p. 1714.

Sacchetta, L. et al., 2022. Synergistic effect of chronic kidney disease, neuropathy, and retinopathy on all-cause mortality in type 1 and type 2 diabetes: a 21-year longitudinal study. *Cardiovascular Diabetology*, 21(1), pp.1-10.

Saha, A., Rajak, S., Saha, J. & Chowdhury, C., 2022. A Survey of Machine Learning and Meta-heuristics Approaches for Sensor-based Human Activity Recognition Systems. *Journal of Ambient Intelligence and Humanized Computing*, pp.1-28.

Sangwung, P., Petersen, K.F., Shulman, G.I. & Knowles, J.W., 2020. Mitochondrial Dysfunction, Insulin Resistance, and Potential Genetic Implications: Potential Role of Alterations in Mitochondrial Function in the Pathogenesis of Insulin Resistance and Type 2 Diabetes. *Endocrinology*, 161(4).

Saravanan, P. et al., 2020. Gestational diabetes: opportunities for improving maternal and child health. *The Lancet Diabetes & Endocrinology*, 8(9), pp.793-800.

Schwarzhans, F. et al., 2024. Intensity Normalization Techniques and Their Effect on the Robustness and Predictive Power of Breast MRI Radiomics. *arXiv preprint arXiv:2406.01736*.

Shaban, M. et al., 2020. A convolutional neural network for the screening and staging of diabetic retinopathy. *Plos One*, 15(6), e0233514.

Shahin, O.R., Alshammari, H.H., Alzahrani, A.A., Alkhiri, H. & Taloba, A.I., 2023. A robust deep neural network framework for the detection of diabetes. *Alexandria Engineering Journal*, 74, pp.715-724.

Shao, H., Liu, X., Zong, D. & Song, Q., 2024. Optimization of diabetes prediction methods based on combinatorial balancing algorithm. *Nutrition & Diabetes*, 14(1), p. 63.

Sharma, P., Hajam, Y.A., Kumar, R. & Rai, S., 2022. Complementary and alternative medicine for the treatment of diabetes and associated complications: A review on therapeutic role of polyphenols. *Phytomedicine Plus*, 2(1), p.100188.

Shrinivasan, L., Verma, R. & Nandeesh, M.D., 2023. Early prediction of diabetes diagnosis using hybrid classification techniques. *IAES International Journal of Artificial Intelligence*, 12(3), pp. 1139–1148.

Sinclair, A.J. & Abdelhafiz, A.H., 2023. Metabolic impact of frailty changes diabetes trajectory. *Metabolites*, 13(2), p.295.

Stephen, O., Sain, M., Maduh, U.J. & Jeong, D.U., 2019. An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare. *Journal of Healthcare Engineering*, 2019, Article ID 4180949, 7 pages. DOI: https://doi.org/10.1155/2019/4180949.

Subbuthai, P. & Muruganand, S., 2015. Restoration of retina images using extended median filter algorithm. *International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE.

Swapna, G., Vinayakumar, R. & Soman, K.P., 2018. Diabetes detection using deep learning algorithms. *ICT Express*, 4(4), pp.243-246.

Thompson, C. & Higgins, O., 2024. Combination of artificial neural network and particle swarm intelligence algorithm for diagnosing diabetes. *Advances in Engineering and Intelligence Systems*, 3(01), pp. 23–33.

Tongur, V. and Ülker, E., 2019. PSO-based improved multi-flocks migrating birds optimization (IMFMBO) algorithm for solution of discrete problems. *Soft Computing*, *23*(14), pp.5469-5484.

Tuncer, T., Dogan, S., Özyurt, F., Belhaouari, S.B. & Bensmail, H., 2020. Novel Multi Center and Threshold Ternary Pattern Based Method for Disease Detection Method Using Voice. *IEEE Access*, 8, pp.84532-84540. DOI: 10.1109/ACCESS.2020.2992641.

Uddin, M.A., Islam, M.M., Talukder, M.A., Hossain, M.A.A., Akhter, A., Aryal, S. & Muntaha, M., 2023. Machine Learning Based Diabetes Detection Model for False Negative Reduction. *Biomedical Materials & Devices*, pp.1-17.

Ulutas, H., Günay, R.B. and Sahin, M.E., 2024. Detecting diabetes in an ensemble model using a unique PSO-GWO hybrid approach to hyperparameter optimization. *Neural Computing and Applications*, pp.1-29.

Usman, T.M., Saheed, Y.K., Ignace, D. & Nsang, A., 2023. Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification. *International Journal of Cognitive Computing in Engineering*, 4, pp.78-88.

Varma, K.M. and Panda, B.S., 2019. Comparative analysis of predicting diabetes using machine learning techniques. *J. Emerg. Technol. Innov. Res*, *6*(6), pp.522-530.

Wang, W., Tong, M. & Yu, M., 2020. Blood Glucose Prediction With VMD and LSTM Optimized by Improved Particle Swarm Optimization. *IEEE Access*, 8, pp.217908-217916. DOI: 10.1109/ACCESS.2020.3041355.

Wheeler, D.C., Stefánsson, B.V., Jongs, N., Chertow, G.M., Greene, T., Hou, F.F., McMurray, J.J.V. et al., 2021. Effects of dapagliflozin on major adverse kidney and cardiovascular events in patients with diabetic and non-diabetic chronic kidney disease: a prespecified analysis from the DAPA-CKD trial. *The Lancet Diabetes & Endocrinology*, 9(1), pp.22-31.

Wicaksana, A.L., Hertanti, N.S., Ferdiana, A. & Pramono, R.B., 2020. Diabetes management and specific considerations for patients with diabetes during coronavirus diseases pandemic: A scoping review. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(5), pp.1109-1120.

Xie, R. (2020, December). A flock-of-starling optimization algorithm with reinforcement learning capability. In *MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (pp. 245-251).

Xu, F., Foong, L.K. & Lyu, Z., 2020. A novel search scheme based on the social behaviour of crow flock for feed-forward learning improvement in predicting the soil compression coefficient. *Engineering with Computers*. DOI: https://doi.org/10.1007/s00366-020-01119-3.

Yan, F., Huang, H., Pedrycz, W. and Hirota, K., 2024. A disease diagnosis system for smart healthcare based on fuzzy clustering and battle royale optimization. *Applied Soft Computing*, *151*, p.111123.

Yang, K., Wang, Y., Li, Y.W., Chen, Y.G., Xing, N., Lin, H.B. & Yu, X.P., 2022. Progress in the treatment of diabetic peripheral neuropathy. *Biomedicine & Pharmacotherapy*, 148, p.112717.

Yin, H., Mukadam, B., Dai, X. & Jha, N.K., 2021. DiabDeep: Pervasive Diabetes Diagnosis Based on Wearable Medical Sensors and Efficient Neural Networks. *IEEE Transactions on Emerging Topics in Computing*, 9(3), pp.1139-1150. DOI: 10.1109/TETC.2019.2958946.

Zak, M. & Krzyżak, A., 2020. Classification of Lung Diseases Using Deep Learning Models. In Krzhizhanovskaya, V.V. et al. (eds) *Computational Science – ICCS 2020*. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-50420-5_47.

Zhang, X. et al., 2020. Deep Learning Based Analysis of Breast Cancer Using Advanced Ensemble Classifier and Linear Discriminant Analysis. *IEEE Access*, 8, pp.120208-120217. DOI: 10.1109/ACCESS.2020.3005228.

Zhang, Y., Wang, S., Sui, Y., Yang, M., Liu, B., Cheng, H., Sun, J., Jia, W., Phillips, P. & Gorriz, J.M., 2018. Multivariate approach for Alzheimer's disease detection using stationary wavelet entropy and predator-prey particle swarm optimization. *Journal of Alzheimer's Disease*, 65(3), pp.855-869.

Zhu, T., Uduku, C., Li, K., Herrero, P., Oliver, N. & Georgiou, P., 2022. Enhancing self-management in type 1 diabetes with wearables and deep learning. *npj Digital Medicine*, 5(1), p.78.

BIODATA OF STUDENT

Balasubramaniyan Divager received his Bachelors degree (B.E) and Masters (M.E) in Computer Science and engineering from Anna University, Tamilnadu, India. He is currently pursuing Doctorate of philosophy (PhD) at the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM). His area of research is intelligent computing. He Contributed to publish peer reviewed journal article and international conference for his research. His area of interest is in the development of Multi Health care related problems and prediction of diseases using the Artificial intelligence techniques such as deep learning models and machine learning models. He is also interested in creating AI based models to treat different health care problems and diseases to scrutinise for the betterment of human welfare and development.

LIST OF PUBLICATIONS

Journal

Divager Balasubramaniyan, Nor Azura Husin, Norwati Mustapha, Nurfadhlina Mohd Sharef, Teh Noranis Mohd Aris "Flock Optimization Induced Deep Learning for Improved Diabetes Disease Classification." *Expert Systems*, 13 Apr. 2023, https://doi.org/10.1111/exsy.13305.

Balasubramaniyan, D., Husin, N. A., Mustapha, N., Sharef, N. M. & Mohd Aris, T. N. (2024). Flock Optimization Algorithm-Based Deep Learning Model for Diabetic Disease Detection Improvement. *Journal of Computer Science*, *20*(2), 168-180. https://doi.org/10.3844/jcssp.2024.168.180

Conference

Divager Balasubramaniyan, Nor Azura Husin, Norwati Mustapha, Nurfadhlina Mohd Sharef and Teh Noranis Mohd Aris "Diabetic disease prediction from Pima Indian Diabetic Dataset using random multinomial logit approach with the spiral interpretable neural network." 4th International Conference on Computer Assisted System in Health, Education and Sustainable Development (CASH).

Divager Balasubramaniyan, Nor Azura Husin, Norwati Mustapha, Nurfadhlina Mohd Sharef and Teh Noranis Mohd Aris "Systematic Review of Current Techniques for Predicting Diabetic Disease from PIMA Indian Dataset." 4th International Conference on Computer Assisted System in Health, Education and Sustainable Development (CASH)