Diabetes Prediction Empowered with Multi-level Data Fusion and Machine Learning

Ghofran Bassam[1], Amina Rouai[2], Reyaz Ahmad[3], Muhammd Adnan Khan[4]
School of Computing, Skyline University College, Sharjah, United Arab Emirates[1, 2, 4]
Department of General Education, Skyline University College, Sharjah, United Arab Emirates[3]
Riphah School of Computing & Innovation-Faculty of Computing, Riphah International University, Lahore, Pakistan[4]
Department of Software-Faculty of Artificial Intelligence and Software, Gachon University, Seongnam-si, Republic of Korea[4]

***Abstract*—Technology improvements have benefited the medical industry, especially in the area of diabetes prediction. In order to find patterns and risk factors related to diabetes, machine learning and Artificial Intelligence (AI) are vital in the analysis of enormous volumes of data, including medical records, lifestyle variables, and biomarkers. This makes it possible for tailored management and early discovery, which might revolutionize healthcare. This study examines how machine learning algorithms may be used to identify diseases, with an emphasis on diabetes prediction. The Proposed Diabetes Prediction Empowered with Mutli-level Data Fusion and Machine Learning (DPEMDFML) model combines two distinct types of models—the Artificial Neural Network (ANN) and the Support Vector Machine (SVM)—to create a fused machine learning technique. Two separate datasets were utilized for training and testing the model in order to assess its performance. To ensure a thorough evaluation of the model's prediction ability, the datasets were split in two experiments in proportions of 70:30 and 75:25, respectively. The study's findings were encouraging, with the ANN algorithm obtaining a remarkable accuracy of 97.43%. This indicates that the model accurately identified instances of diabetes, indicating a high degree of accuracy. A more thorough knowledge of the model's prediction ability would result from further assessment and validation of its performance using various measures.**

Keywords*—Disease prediction; machine learning (ML); fused approach; artificial neural network (ANN); support vector machine (SVM); disease diagnosis; healthcare

I. Introduction

The chronic metabolic condition known as diabetes affects millions of people worldwide. The World Health Organization projects that by 2030, 643 million people worldwide will have diabetes, up from an expected 537 million in 2021 [1]. Diabetes is brought on by abnormalities in insulin synthesis or function, which hinder the body from effectively managing blood sugar levels. All ages are impacted, and if it is not treated, it might have detrimental implications on one's health. The body's immune system wrongly assaults and destroys pancreatic insulin-producing cells in autoimmune type 1 diabetes [2]. It usually appears during childhood or adolescence and necessitates lifelong insulin medication. Obesity, inactivity, and poor eating habits are commonly linked to the majority of type 2 diabetes cases [3]. Type 2 diabetes is differentiated by a decrease in the body's ability to produce enough insulin to maintain normal blood sugar levels or by an increase in insulin resistance [3]. Numerous consequences can result from unmanaged diabetes. For diabetics, cardiovascular disease, such as heart attacks and strokes, is a major worry. Kidney issues, nerve damage (neuropathy), retinopathy, and foot issues are some of the consequences of diabetes [4]. One's quality of life may be significantly impacted by these problems, which need continual medical care. Traditional diabetes prediction systems confront a number of problems. These techniques frequently depend on simplistic statistical models or rudimentary machine learning algorithms, which are incapable of capturing the intricate interplay of many risk variables. Furthermore, these techniques may underutilize the potential of accessible data sources such as patient medical records, genetic information, lifestyle variables, and environmental factors. As a result, the accuracy and reliability of diabetes prediction using these traditional methods are inadequate. A subset of artificial intelligence called machine learning has completely changed several industries, including the healthcare industry. It involves developing algorithms and models that are able to absorb knowledge from data and act or anticipate without being explicitly programmed. The medical sector's decision-making processes for disease prediction, diagnosis, and treatment have showed great promise when using machine learning techniques. Researchers have investigated the merging of different ML methods for diabetes prediction in order to overcome the limitations of existing methodologies (Table I). Fusing several algorithms enables for the use of each method's distinct strengths while correcting for their particular flaws and improving forecast accuracy. A fused machine learning model can give a more thorough and holistic view of the condition by merging diverse data sources such as electronic health records, medical imaging, genetic profiles, and lifestyle data [5]. An ML-based diagnostic system can help detect diabetic patients early on which leads improve patient outcomes and help lessen the burden of diabetes on individuals and healthcare systems. This paper presents a unique framework utilizing machine learning fusion to achieve early diagnosis of diabetes patients. The system goals to increase the accuracy and efficacy of diabetes diagnosis by combining various machine learning algorithms and diverse datasets. This approach leads to proactive healthcare interventions and ultimately improves patient outcomes.

The Proposed Diabetes Prediction Empowered with Mutli-level Data Fusion and Machine Learning (DPEMDFML) model framework is presenting diabetes disease prediction. It is carried out using the ANN and SVM algorithms, while using two different datasets.

The IoMT is necessary for enhancing the accuracy, reliability, and efficacy of electronic equipment in the medical field. By integrating the existing health care assets and medical facilities, experts are advancing a digital medical system [6]. The control of infectious disease waves is eased by prompt diagnosis and improved ongoing treatment. The internet of medical things (IoMT) is a growing area of technology that is now being used to assist Point-of-care testing (POCT). Using the IoMT, POCT devices may operate wirelessly and be connected to health professionals and medical facilities [7].

Recently has been discovered that developed ANNs may perform well in a variety of circumstances due to ANNs' universal prediction capabilities and adaptable network architectures [8]. The building block of the ANN created to mimic the function of a human neuron. Also, one of the greatest methods for analyzing data is the use of SVM. To control data, they utilize generalization controlling [9]. SVM is an artificial intelligence method that assigns labels to things by learning from examples [10]. The innovative and promising IoMT framework presented in this study represents a significant leap forward in the realm of diabetes disease prediction. Drawing upon the capabilities of two cutting-edge machine learning algorithms, ANN and SVM, this framework exemplifies the fusion of advanced technology and healthcare, offering a transformative approach to diabetes management and patient care. At its core, the IoMT framework capitalizes on the vast amount of data generated by interconnected medical devices, wearable sensors, and health monitoring systems. By harnessing this continuous and diverse stream of patient-specific information, healthcare providers gain unprecedented insights into the multifaceted aspects of diabetes, allowing for more precise, proactive, and personalized interventions. The first pillar of the framework, Artificial Neural Networks (ANN), represents a sophisticated computational model inspired by the complex interconnections of neurons in the human brain. ANN's ability to learn from data and recognize intricate patterns and non-linear relationships makes it an ideal candidate for diabetes prediction. The network's architecture is meticulously designed, leveraging multiple layers of interconnected neurons to extract high-level features from raw input data. The ANN's adaptability enables it to adjust its internal parameters during the learning process, optimizing the model's performance to achieve highly accurate diabetes predictions. In tandem with ANN, the IoMT framework also incorporates the renowned Support Vector Machine (SVM) algorithm, renowned for its prowess in binary classification tasks and its ability to handle complex decision boundaries. SVM's kernel-based approach allows it to efficiently discover non-linear patterns in the feature space, making it invaluable for diabetes prediction when the relationship between features and disease occurrence is intricate and not easily separable.

By integrating the capabilities of both ANN and SVM, the IoMT framework achieves a powerful ensemble of predictive models that complement each other's strengths. The diversity of these algorithms enhances the framework's ability to capture subtle nuances and intricate interactions within the data, ultimately leading to more reliable and accurate diabetes predictions. Data privacy and security are of paramount concern within the IoMT framework. Stringent measures are implemented to anonymize and safeguard patient information, and access controls are enforced to protect sensitive data from unauthorized disclosure. The framework's design ensures that data is utilized solely for model training purposes, mitigating the risk of data breaches and preserving patient confidentiality. The synergistic integration of ANNs and SVM algorithms within the IoMT framework marks a significant step towards personalized and data-driven diabetes prediction. With the potential to revolutionize healthcare practices, this cutting-edge approach empowers clinicians with actionable insights, fosters early detection, and facilitates effective diabetes management, ultimately enhancing the quality of life for patients worldwide.

The structure of the research paper is as follows: Section II represents the related work. In Section III, the contribution is presented. The detail of the proposed model is described in Section IV. Discussion and analysis of results are discussed in Section V. The conclusion of this research is presented in Section VI.

II. Related Work

The presented findings encompass various studies that examined different healthcare databases and utilized diverse approaches and strategies to make predictions. Researchers have developed and employed a range of prediction models, incorporating various data mining techniques, algorithmic methods for machine learning, or even a combination of these strategies. These studies highlight the wide array of approaches utilized in healthcare research to enhance prediction accuracy and improve decision-making processes.

Akkarapol and Jongsawas [11] presented a paper that analysed a dataset comprising 50,788 records with 43 parameters. The research identified significant risk variables, including age, BMI, overall revenue, sex, heart attack history, marital status, dentist check-up frequency, and diagnosis of asthma. Other risk factors such as hypertension and cholesterol were also recognized. The study's overall reliability was reported as 77.11%, indicating a moderate level of consistency in the findings. Furthermore, the true negative rate specifically for the Artificial Neural Network (ANN) model was noted as 79.45%, indicating its ability to accurately identify negative cases.

Kavakiotis et al.'s paper [12] focused on evaluating data mining and machine learning techniques for DM research. Through the systematic comparison of three algorithms, including Logistic Regression, Naive Bayes, and SVM, using 10-fold cross-validation, the study concluded that SVM achieved the highest accuracy rate of 84%. These findings contribute to the understanding of algorithm selection in DM research, highlighting the potential benefits of SVM in achieving accurate predictions and improving decision-making processes.

Xue-Hui Meng et al.'s study [13] focused on comparing the performance of decision tree models, ANNs, and logistic regression in diagnosing diabetes or prediabetes based on general risk variables. The logistic regression model achieved a classification accuracy of 76.13%, indicating its ability to correctly classify individuals as having diabetes or prediabetes based on the general risk variables considered in the study. The decision tree model (C5.0) demonstrated a slightly higher classification accuracy of 77.87%. It also showed a relatively high sensitivity of 80.68%, meaning it successfully identified a large proportion of True Positive (TP) cases, and a specificity of 75.13%, indicating its capability to accurately identify True Negative (TN) cases. In contrast, the ANN model obtained a lower classification accuracy of 73.23%, suggesting that it was less effective in predicting the disease outcomes using the same set of general risk variables.

The research work conducted by Md. Faisal Faruque, Asaduzzaman, and Iqbal [14] focused on exploring the relationship between Diabetes Mellitus and multiple risk factors through the analysis of 16 attributes including factors such as age, diet, hypertension, vision problems, and genetic predisposition. By utilizing four popular machine learning algorithms, the researchers examined data from 200 patients. The findings of the study indicated that the Decision Tree algorithm demonstrated superior predictive performance compared to Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbour (KNN) algorithms in this particular study, suggesting its potential efficacy in predicting or classifying the disease based on the identified risk factors.

TABLE I. LIMITATIONS OF THE PREVIOUS WORKS

Research Study	Method	Accuracy	Limitation
Akkarapol and Jongsawas [11]		77.11%	- Low accuracy - Limited to a specific region
Kavakiotis et al. [12]		84%	- Used three algorithms but the accuracy is low.
Xue-Hui Meng et al. [13]	Logistic Regression Model Decision Tree Model (C5.0) Artificial Neural Networks (ANN) Model	76.13% 77.87% 73.23%	- Low accuracy - Limited features of the dataset used
Md. Faisal Faruque, Asaduzzaman, and Iqbal [14]	Decision Tree Algorithm Support Vector Machine (SVM) Naive Bayes (NB) K-Nearest Neighbour (KNN)	Not specified	- Small sample size - Limited features
Dey et al. [15]	ANN Model with MMS	82.35%	- Limited evaluation matrices
Pradhan et al. [16]	Ensemble Learning Approach	Not specified	- Multiple algorithms without mentioning accuracies

The study conducted by Dey et al. [15] utilized four well-known supervised machine learning algorithms: SVM, KNN, Naive Bayes, and ANN with MMS. These algorithms were selected for their ability to learn from labelled data and make predictions based on learned patterns and relationships to analyse the Pima Indian dataset. The study revealed that the ANN model with MMS achieved the highest accuracy rate of 82.35%, indicating its potential effectiveness in predicting the specific outcome compared to the other four algorithms examined.

Pradhan et al. research [16] employed supervised learning, which involves training models on labelled data to make predictions, to develop models for diabetes diagnosis. Additionally, they utilized hybrid learning, which combines multiple learning techniques, to further enhance the performance of the diagnostic models. Finally, the researchers explored ensemble learning, a powerful approach that combines the predictions of multiple individual models, to create a more robust and accurate diabetes diagnosis model. The results of the study demonstrated that the ensemble learning approach surpassed both supervised learning and hybrid learning in terms of accuracy.

III. CONTRIBUTION

In contrast to previous research, this Diabetes Prediction Empowered with Multi-level Data Fusion and Machine Learning (DPEMDFML) model represents a more comprehensive study that explores various commonly used techniques for diabetes identification. The primary objective is to compare the performance of these techniques and identify the most effective one. It has been accomplished by employing two distinct algorithms and evaluating them on two different datasets, considering all relevant evaluation metrics. Furthermore, this study delves into analyzing the significance of each attribute in influencing the classification outcome. This analysis provides valuable insights for future research to adapt and improve the dataset, making it more informative and suitable for diabetes diagnosis tasks.

IV. PROPOSED MODEL

The Diabetes Prediction Empowered with Multi-level Data Fusion and Machine Learning (DPEMDFML) model developed here seeks to predict diabetes in a smart healthcare system utilizing data from the Internet of Medical Things (IoMT) is divided into two stages: training and testing as shown in Fig. 1. During the Training Phase, hospitals (Hospitals A, B, C, and N) use IoMT devices to gather patient data, which is subsequently recorded in their respective local databases. This information might include vital indicators, blood glucose levels, lifestyle information, and other information. The 'Prediction Layer,' which houses multiple ML models, with a focus on Support Vector Machines (SVM) and Artificial Neural Networks (ANN), is at the core of this phase.

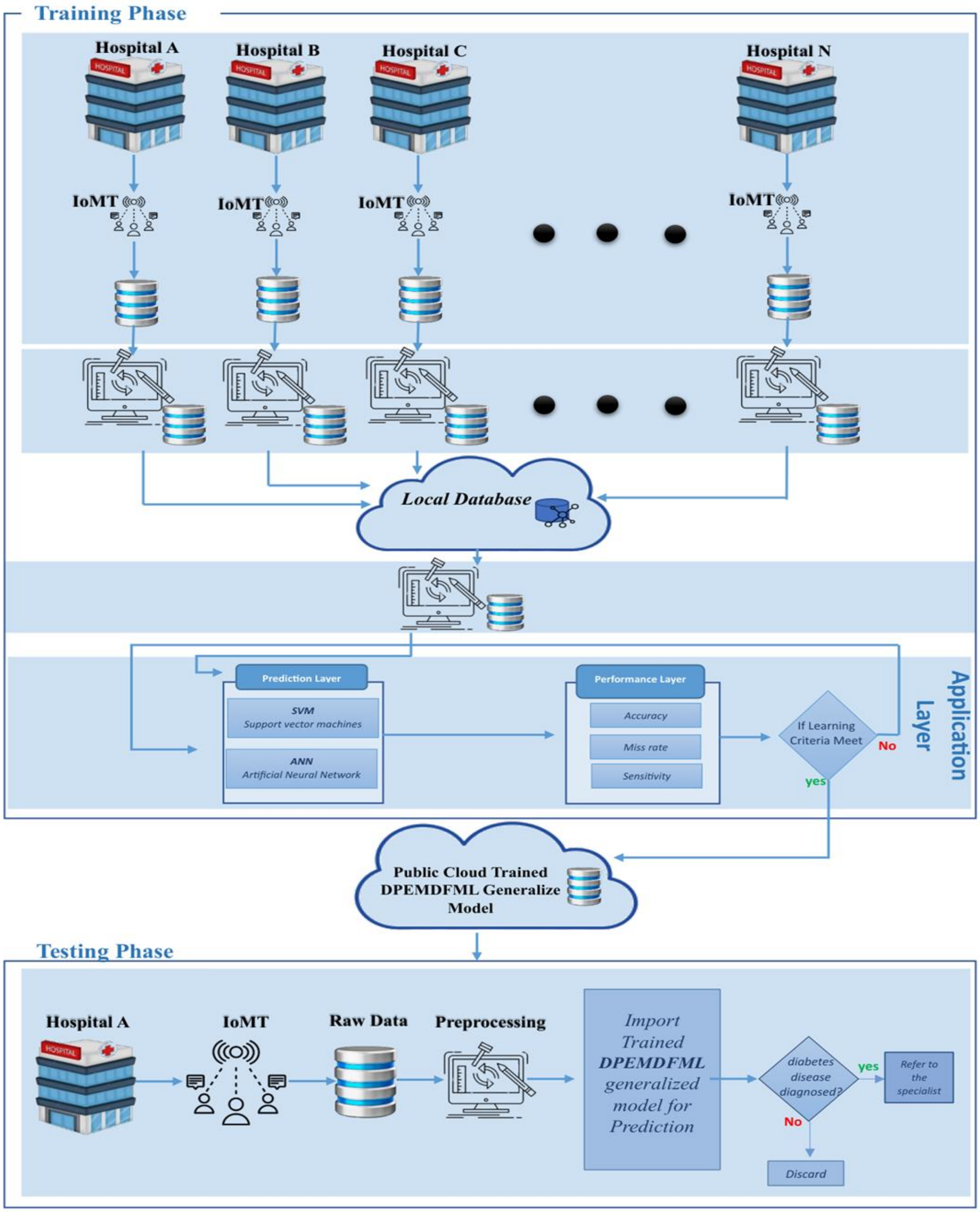

Fig. 1. Diabetes prediction empowered with multi-level data fusion and machine learning (DPEMDFML).

These models excel at classification tasks and are in charge of learning whether a patient has diabetes depending on the input data. Following the Prediction Layer, Fig. 1 shows the Performance Layer assesses the efficiency of the ML models by employing measures such as accuracy, miss rate, and sensitivity. Models that match the performance criteria are saved in the public cloud as the "DPEMDFML Generalized Model", while those that fall short go through additional training rounds to enhance accuracy.

The trained DPEMDFML Generalized Model is used in the Testing Phase. When new patients from Hospital N seek diabetes diagnosis, the system gathers raw data from IoMT devices, which is then processed. Data cleansing, value normalization, and missing data management are examples of pre-processing operations that ensure the input data is ideal for the ML models' predictions.

The DPEMDFML Generalized Model is then used to predict whether or not the patient has diabetes. This decision-making procedure has two results: If diabetes is predicted by the model, the patient is directed to a specialist for prompt medical intervention. If the model predicts a poor outcome, the data is properly deleted, protecting patient confidentiality and data privacy.

Because the system is distributed, various hospitals can contribute data, resulting in a broad and complete dataset for model training. Furthermore, the cloud-based architecture improves accessibility and scalability, allowing the system to meet growing data volumes as well as changing healthcare demands. The system benefits from the capabilities of SVM and ANN as its major ML models in pattern recognition, feature extraction, and classification, results in accurate diabetes predictions. Furthermore, the system's iterative training technique allows for continuous development, keeping the models current with medical advances.

The relevance of this ML-driven approach resides in its potential to improve diabetes diagnosis and patient treatment. The approach leverages the available information by utilizing data from IoMT devices across many hospitals, resulting in more reliable and exact predictions. The capacity to detect diabetic patients quickly and give early medical treatment assures improved disease control and perhaps improves patient outcomes. As the system evolves, its influence on the healthcare environment is expected to go beyond diabetes diagnosis, with the ability to tackle additional medical difficulties utilizing a similar distributed, ML-based approach.

The distributed, cloud-based machine learning system for diabetes detection using IoMT data is a potential improvement in healthcare technology. Its training and testing phases, which are supported by SVM and ANN models, show that it can handle complicated medical data and make correct predictions. As the system evolves via iterative training and embraces an ever-growing dataset, it is positioned to impact the future of medical diagnosis, eventually improving patient care and contributing to the healthcare industry's continuing transformation.

A. Datasets

Diabetes Prediction Empowered with Mutli-level Data Fusion and Machine Learning (DPEMDFML) Model used two different datasets:

The primary dataset employed in this research is the PIMA Indian Diabetes Database, accessible at the University of California machine learning repository [14]. The dataset encompasses information from 768 individuals, all of whom are female, and their ages span from 21 to 81 years. For each individual, the dataset consists of nine distinct feature characteristics. These feature characteristics include eight continuous quantitative variables, namely the number of pregnancies, blood sugar level (in mg/dL), diastolic blood pressure (in mmHg), skin fold thickness (in mm), body mass index (BMI), serum insulin level (in mU/mL), age (in years), and a pedigree function associated with diabetes. By utilizing this comprehensive dataset, the study aims to explore the relationships between these feature characteristics and diabetes occurrence, enabling the development of predictive models for early detection and assessment of diabetes risk in female patients.

For the second dataset used in this paper, it is called the "Diabetes prediction dataset," sourced from Electronic Health Records (EHRs) [15]. The dataset encompasses information from a substantial sample of 100,000 individuals, which were collected from diverse healthcare providers and then aggregated into a unified dataset. It is noteworthy that this dataset includes both female and male participants. The Diabetes prediction dataset consists of eight distinctive feature characteristics for each individual. These features include age, gender, hypertension, heart disease, smoking history, BMI (body mass index), HBA1C level (glycated haemoglobin level), and glucose level. By utilizing this comprehensive dataset, the study aims to explore the relationships between these feature characteristics and diabetes prediction. The inclusion of both genders and the diverse range of feature characteristics in this dataset facilitate a comprehensive analysis, providing valuable insights into predicting diabetes and its associated risk factors.

V. Results and Discussion

This section showcases the results of diabetes prediction using two different machine learning models: Support Vector Machine (SVM) and Artificial Neural Network (ANN). The prediction is conducted on two distinct datasets, and each dataset is split into two different ratios for training and testing: 70:30 and 75:25. Then, a range of evaluation metrics are calculated, include accuracy, miss-classification rate, sensitivity, specificity, precision, False positive (FP) rate, False discovery rate, False omission rate, Positive likelihood ratio,

Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio. The following equations illustrate the equations used to calculate each of these metrics, providing a clear understanding of the underlying mathematical formulas for the statistical measurements [17-23]. The utilization of this diverse set of metrics ensures a comprehensive assessment of the models’ performance, accounting for different aspects of predictive accuracy and error rates. Python is utilized as the simulation tool for implementing both the SVM model and ANN model, to obtain the results.

$$\text{Accuracy} = \frac{\text{Total Number of Predictions}}{\text{Number of Correct Predictions}} \tag{1}$$

$$\text{Miss classification rate } = 1 - \text{Accuracy} \tag{2}$$

$$Sensitivity = \frac{\text{True Positive}}{\text{True Positives+False Negatives}} \tag{3}$$

$$Specificity = \frac{\text{True Negatives}}{\text{True Negatives+False Positives}} \tag{4}$$

$$Precision = \frac{\text{True Positives}}{\text{True Positives+True Positives}} \tag{5}$$

$$False\ positive\ rate = \frac{\text{False Positives}}{\text{False Positives+True Negatives}} \tag{6}$$

$$False\ discovery\ rate = \frac{\text{False Positives}}{\text{False Positives+True Positives}} \tag{7}$$

$$False\ omission\ rate = \frac{\text{False Negatives}}{\text{False Negatives+True Negatives}} \quad (8)$$

$$Positive\ likelihood\ ration = \frac{\text{Sensitivity}}{1-\text{Specificity}} \quad (9)$$

$$Negative\ likelihood\ ration = \frac{1-\text{ Sensitivity}}{\text{Specificity}} \quad (10)$$

$$Prevalence\ thresholde = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (11)$$

$$Critical\ sucess\ index = \frac{\text{True Positives}}{\text{True Positives+False Negatives+False Positives}} \quad (12)$$

$$F1\ score = \frac{2\times\text{Precision}\times\text{Sensitivity}}{\text{Precision+Sensitivity}} \quad (13)$$

$$MCC = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \quad (14)$$

$$FMI = \sqrt{\text{Sensitivity} \times \text{Precision}} \quad (15)$$

$$Informedness = \text{ Sensitivity} + \text{Specificity} - 1 (16)$$

$$Diagnostic\ odds\ ratio = \frac{\text{Positive likelihood ration}}{\text{Negative likelihood ration}} \quad (17)$$

A. DPEMDFML - SVM System Model - using Pima Diabetes Dataset - 70:30

Using SVM model with the Pima Diabetes Dataset, the dataset is divided as: 30% for testing (n=231) and 70% for training (n=537) to assess the model's performance accurately. The performance evaluation of the SVM model is depicted in Table II and Table III, which illustrate the confusion matrix. The confusion matrix provides crucial insights into the model's predictive accuracy, enabling a detailed examination of how well the SVM algorithm classifies diabetes and non-diabetes cases in the dataset.

During the training phase, the SVM model’s predictions for diabetes disease are presented in Table II. The training dataset consists of 537 samples, which are further categorized into 187 real positive samples, indicating the presence of diabetes, and 350 real negative samples, indicating the absence of diabetes. Among the real positive samples (indicating the presence of diabetes), the SVM model correctly identifies 117 samples as positive, accurately signaling the presence of healthcare issues. However, the model misclassifies 70 records as negatives, incorrectly suggesting the absence of healthcare issues when there is an actual health concern. On the other hand, among the real negative samples (indicating the absence of diabetes), the SVM model correctly predicts 309 samples as negative, appropriately identifying the absence of healthcare conditions. However, the model makes errors in 41 samples, wrongly classifying them as positive, inaccurately indicating the presence of a healthcare issue when there is none.

During the testing phase, the SVM model's predictions for diabetes disease are presented in Table III. The testing dataset consists of 231 samples, which are further categorized into 81 real positive samples, indicating the presence of diabetes, and 150 real negative samples, indicating the absence of diabetes. Among the real positive samples (indicating the presence of diabetes), the SVM model correctly identifies 48 samples as positive, accurately signaling the presence of healthcare issues. However, the model misclassifies 33 records as negatives, incorrectly suggesting the absence of healthcare issues when there is an actual health concern. However, the SVM model successfully predicted 124 samples as negative, properly recognizing the lack of medical diseases among the genuine negative samples (showing the absence of diabetes). But in 26 samples, the model misclassifies them as positive, thus implying the existence of a healthcare concern when there isn't one.

Table IV presents a comprehensive overview of the performance of the proposed SVM model in terms of various evaluation metrics. During the training phase, the SVM model achieved the following percentages for each metric: 79.32% accuracy, 20.67% miss-classification rate, 62.56% sensitivity, 88.28% specificity, 74.05% precision, 11.71% False positive rate, 25.94% False discovery rate, 18.46% False omission rate, 534.10% Positive likelihood ratio, 478.00% Negative likelihood ratio, 30.20% Prevalence threshold, 51.31% critical success index, 67.82% F1 Score, 53.16% Mathews Correlation coefficient, 68.06% Fowlkes-Mallows Index, 50.85% informedness, and 1259.68% Diagnostic odds ratio. During the validation phase, the performance of the model is evaluated, and the following evaluation metrics are obtained: 74.46% accuracy, 25.54% miss-classification rate, 59.25% sensitivity, 82.66% specificity, 64.86% precision, 17.33% False positive rate, 35.13% False discovery rate, 21.01% False omission rate, 341.88% Positive likelihood ratio, 393.29% Negative likelihood ratio, 35.10% Prevalence threshold, 44.86% critical success index, 61.93% F1 Score, 42.87% Mathews Correlation coefficient, 61.99% Fowlkes-Mallows Index, 41.92% informedness, and 693.70% Diagnostic odds ratio.

TABLE II. SVM MODEL'S: PIMA DIABETES DATASET – TRAINING PHASE – 70:30

Input	**Total number of samples (537)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	187(positive)	117(TP)	70(FN)
	350(negative)	41(FP)	309(TN)

TABLE III. SVM MODEL'S: PIMA DIABETES DATASET – TESTING PHASE – 70:30

Input	**Total number of samples (231)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	81(positive)	48(TP)	33(FN)
	150(negative)	26(FP)	124(TN)

TABLE IV. SVM Model's (Pima Diabetes Dataset) Evaluation Metrics, 70:30

	Testing	**Training**
Accuracy	0.7445 (74.46 %)	0.7932 (79.32 %)
Miss-classification rate	0.2554 (25.54 %)	0.2067 (20.67 %)
Sensitivity	0.5925 (59.25 %)	0.6256 (62.56 %)
Specificity	0.8266 (82.66 %)	0.8828 (88.28 %)
Precision	0.6486 (64.86 %)	0.7405 (74.05 %)
False positive rate	0.1733 (17.33 %)	0.1171 (11.71 %)
False discovery rate	0.3513 (35.13 %)	0.2594 (25.94%)
false omission rate	0.2101 (21.01 %)	0.1846 (18.46 %)
Positive likelihood ration	3.4188 (341.88 %)	5.3410 (534.10 %)
Negative likelihood ratio	3.9329 (393.29 %)	4.7800 (478.00 %)
prevalence threshold	0.3510 (35.10 %)	0.3020 (30.20 %)
critical success index	0.4485 (44.859 %)	0.5131 (51.31 %)
F1 Score	0.6193 (61.93 %)	0.6782 (67.82 %)
Mathews Correlation co-efficient	0.4287 (42.87 %)	0.5316 (53.16 %)
Fowlkes-Mallows Index	0.6199 (61.99 %)	0.6806 (68.06 %)
informedness	0.4192 (41.92 %)	0.5085 (50.85 %)
Diagnostic odds ratio	6.9370 (693.70 %)	12.5968 (1259.68 %)

B. DPEMDFML - SVM System Model - using Pima Diabetes Dataset - 75:25

Again, using SVM model with the Pima Diabetes Dataset. The dataset is divided into 25% for testing (n=192) and 75% for training (n=576) to assess the model's performance accurately. The performance evaluation of the SVM model is depicted in Table V and Table VI, which illustrate the confusion matrix.

Table V demonstrates the performance of the SVM model in predicting diabetic illness during the training phase. The training dataset comprises 576 samples, with 203 being true positive cases, indicating the presence of diabetes, and 373 being true negative cases, indicating the absence of diabetes. For the true positive cases, the SVM algorithm successfully identifies and correctly classifies 124 samples as positive, meaning that it accurately detects the absence of healthcare problems in those cases. However, the algorithm makes 79 errors by misclassifying some samples as negatives, falsely suggesting the absence of healthcare concerns when diabetes is actually present. Regarding the true negative cases, the SVM model performs well by accurately predicting and classifying 330 samples as negative, properly recognizing the absence of diabetes and the presence of other medical issues in those cases. Nevertheless, the model misclassifies 43 samples as positive, falsely indicating the presence of a healthcare issue when there is, in fact, no such health concern.

TABLE V. SVM MODEL'S - PIMA DIABETES DATASET – TRAINING PHASE – 75:25

Input	Total number of samples (576)	Result (output)	
	Expected output	Predicted positive	Predicted negative
	203(positive)	*124(TP)*	*79(FN)*
	373 (negative)	*43(FP)*	*330(TN)*

During the testing phase, Table VI showcases the SVM model's predictions for diabetes disease. The testing dataset consists of 192 samples, which are further categorized into 65 real positive samples, indicating the presence of diabetes, and 127 real negative samples, indicating the absence of diabetes. Among the real positive samples (indicating the presence of diabetes), the SVM model correctly identifies 36 samples as positive, accurately signaling the absence of healthcare issues. However, the model misclassifies 29 records as negatives, incorrectly suggesting the presence of healthcare issues when there is none. On the other hand, among the real negative samples (indicating the absence of diabetes), the SVM model correctly predicts 105 samples as negative, appropriately identifying the presence of healthcare conditions. However, the model makes errors in 22 samples, wrongly classifying them as positive, inaccurately indicating the absence of a healthcare issue when there is a health concern.

Table VII presents a comprehensive overview of the performance of the proposed SVM model in terms of various evaluation metrics. During the training phase, the SVM model achieved the following percentages for each metric: 78.81%, 21.18%, 61.08%, 88.47%, 74.25%, 11.52%, 25.74%, 19.31 %, 529.86 %, 458.03 %, 30.82%, 50.40%, 67.02%, 52.17%, 67.34%, 49.55%, 1204.59%, accuracy, miss-classification rate, sensitivity, specificity, precision, False positive rate, False discovery rate, False omission rate, Positive likelihood ratio, Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio, respectively. During the validation phase, the performance of the model is evaluated, and the following evaluation metrics are obtained: 73.43% accuracy, 26.56 % miss-classification rate, 55. 38 % sensitivity, 82.67% specificity, 62.06% precision, 17.32% False positive rate, 37. 93% False discovery rate, 21.64% False omission rate, 319.72% Positive likelihood ratio, 382.02% Negative likelihood ratio, 35.86% Prevalence threshold, 41.37% critical success index, 58.53% F1 Score, 39.22% Mathews Correlation coefficient, 58.63% Fowlkes-Mallows Index, 38.06% informedness, and 592.47% Diagnostic odds ratio.

TABLE VI. SVM MODEL'S - PIMA DIABETES DATASET – TESTING PHASE – 75:25

	Testing	Training
Accuracy	0.7343 (73.43 %)	0.7881 (78.81 %)
Miss-classification rate	0.2656 (26.56 %)	0.2118 (21.18 %)
Sensitivity	0.5538 (55.38 %)	0.6108 (61.08 %)
Specificity	0.8267 (82.67 %)	0.8847 (88.47 %)
Precision	0.6206 (62.06 %)	0.7425 (74.25 %)
False positive rate	0.1732 (17.32 %)	0.1152 (11.52 %)
False discovery rate	0.3793 (37. 93 %)	0.2574 (25.74%)
false omission rate	0.2164 (21.64%)	0.1931 (19.31 %)
Positive likelihood ration	3.1972 (319.72 %)	5.2986 (529.86 %)
Negative likelihood ratio	3.8202 (382.02 %)	4.5803 (458.03 %)
prevalence threshold	0.3586 (35.86 %)	0.3028 (30.82 %)
critical success index	0.4137 (41.37 %)	0.5040 (50.40 %)

F1 Score	0.5853 (58.53 %)	0.6702 (67.02 %)
Mathews Correlation co-efficient	0.3922 (39.22 %)	0.5217 (52.17 %)
Fowlkes-Mallows Index	0.5863 (58.63 %)	0.6734 (67.34 %)
informedness	0.3806 (38.06 %)	0.4955 (49.55 %)
Diagnostic odds ratio	5.9247 (592.47 %)	12.0459 (1204.59 %)

TABLE VII. SVM MODEL'S (PIMA DIABETES DATASET) EVALUATION METRICS, 75:25

	Testing	**Training**
Accuracy	0.7343 (73.43 %)	0.7881 (78.81 %)
Miss-classification rate	0.2656 (26.56 %)	0.2118 (21.18 %)
Sensitivity	0.5538 (55.38 %)	0.6108 (61.08 %)
Specificity	0.8267 (82.67 %)	0.8847 (88.47 %)
Precision	0.6206 (62.06 %)	0.7425 (74.25 %)
False positive rate	0.1732 (17.32 %)	0.1152 (11.52 %)
False discovery rate	0.3793 (37. 93 %)	0.2574 (25.74%)
false omission rate	0.2164 (21.64%)	0.1931 (19.31 %)
Positive likelihood ration	3.1972 (319.72 %)	5.2986 (529.86 %)
Negative likelihood ratio	3.8202 (382.02 %)	4.5803 (458.03 %)
prevalence threshold	0.3586 (35.86 %)	0.3028 (30.82 %)
critical success index	0.4137 (41.37 %)	0.5040 (50.40 %)
F1 Score	0.5853 (58.53 %)	0.6702 (67.02 %)
Mathews Correlation co-efficient	0.3922 (39.22 %)	0.5217 (52.17 %)
Fowlkes-Mallows Index	0.5863 (58.63 %)	0.6734 (67.34 %)
informedness	0.3806 (38.06 %)	0.4955 (49.55 %)
Diagnostic odds ratio	5.9247 (592.47 %)	12.0459 (1204.59 %)

C. DPEMDFML - SVM System Model - using EHRs Dataset - 70:30

The SVM model was utilized in this study with the EHRs Dataset (Electronic Health Records Dataset). To ensure a robust evaluation of the model's performance, the dataset was divided into 30% for testing (n=30,000) and 70% for training (n=70,000). To assess the effectiveness of the SVM model, its performance was analysed using two distinct evaluation tables: Table VIII and Table IX, both presenting the confusion matrix.

During the training phase, the SVM model's diabetes predictions are presented in Table VIII. The training dataset consists of an extensive sample of 70,000 records, which are further categorized into 5,972 instances as positive cases, indicating the presence of diabetes, and 64,028 instances as negative cases, indicating the absence of diabetes. Among the actual positive cases, the SVM model correctly identifies 3,621 samples as positive, correctly indicating the absence of healthcare issues. However, the model misclassifies 2,351 records as negative, falsely signalling the presence of healthcare issues where there are none. On the other hand, among the actual negative cases, the SVM model accurately predicts 63,602 samples as negative, correctly identifying the presence of healthcare conditions. However, the model makes errors in 426 samples, incorrectly classifying them as positive, falsely indicating the absence of a healthcare issue.

During the testing phase, the SVM model's predictions for diabetes disease are displayed in Table IX. The testing dataset comprises 30,000 samples, which are further categorized into 2,528 true positive cases, indicating the presence of diabetes, and 27,472 true negative cases, indicating the absence of diabetes. Among the true positive cases, the SVM model correctly classifies 1,515 samples as positive, accurately indicating the absence of any healthcare issues. However, the model misclassifies 1,013 records as negative, falsely indicating the presence of healthcare issues when there are none. Conversely, among the true negative cases, the SVM model accurately predicts 27,298 samples as negative, correctly identifying the presence of healthcare conditions. Nevertheless, the model makes errors in 174 samples, incorrectly classifying them as positive, falsely indicating the absence of a healthcare issue.

Table X presents a comprehensive overview of the performance of the proposed SVM model in terms of various evaluation metrics. During the training phase, the SVM model achieved the following percentages for each metric: 96.03%, 3.96%, 60.63%, 99.33%, 89.47%, 0.66%, 10.52%, 3.56%, 9113.16%, 2786.65%, 9.48%, 56.59%, 72.28%, 71.77%, 73.65%, 59.96%, 22995.19%, accuracy, miss-classification rate, sensitivity, specificity, precision, False positive rate, False discovery rate, False omission rate, Positive likelihood ratio, Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio, respectively. During the validation phase, the performance of the model is evaluated, and the following evaluation metrics are obtained: 96.04%, 3.95%, 59.92%, 99.36%, 89.69%, 0.63%, 10.3%, 3.57%, 9461.86%, 2777.06%, 9.32%, 56.06%, 71.85%, 71.45%, 73.31%, 59.29%, 23463.06%, accuracy, miss-classification rate, sensitivity, specificity, precision, False positive rate, False discovery rate, False omission rate, Positive likelihood ratio, Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio, respectively.

D. DPEMDFML - SVM System Model - using EHRs Dataset – 75:25

Here, the SVM model with the EHRs Dataset (Electronic Health Records Dataset) is employed. To ensure a robust assessment of the model's performance, the dataset was split into 25% for testing (n=25,000) and 75% for training (n=75,000). To evaluate the SVM model's effectiveness, two different evaluation tables was used to analyse its performance: Table XI and Table XII, which present the confusion matrix.

During the training phase, the SVM model's diabetes predictions are presented in Table XI. The training dataset consists of 75,000 samples, which are further categorized into 6,409 true positive cases, indicating the presence of diabetes, and 68,591 true negative cases, indicating the absence of diabetes. Among the true positive cases, the SVM model correctly identifies 3,876 samples as positive, accurately indicating the absence of healthcare issues. However, the model misclassifies 2,533 records as negative, falsely signalling the presence of healthcare issues where there are none. On the other hand, among the true negative cases, the SVM model accurately predicts 68,111 samples as negative, correctly identifying the presence of healthcare conditions. However, the model makes errors in 480 samples, incorrectly classifying them as positive, falsely indicating the absence of a healthcare issue.

TABLE VIII. SVM MODEL'S - EHRS DIABETES DATASET – TRAINING PHASE – 70:30

Input	**Total number of samples (70000)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	5972(positive)	*3621(TP)*	*2351(FN)*
	64028(negative)	*426(FP)*	*63602(TN)*

TABLE IX. SVM MODEL'S - EHRS DIABETES DATASET – TESTING PHASE – 70:30

Input	**Total number of samples (30000)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	2528(positive)	*1515(TP)*	*1013(FN)*
	27472(negative)	*174(FP)*	*27298(TN)*

TABLE X. SVM MODEL'S (EHRS DIABETES DATASET) EVALUATION METRICS, 70:30

	Testing	**Training**
Accuracy	0.9604 (96.04 %)	0.9603 (96.03 %)
Miss-classification rate	0.0395 (3. 95 %)	0.0396 (3.96 %)
Sensitivity	0.5992 (59.92 %)	0.6063 (60.63 %)
Specificity	0.9936 (99.36 %)	0.9933 (99.33 %)
Precision	0.8969 (89.69 %)	0.8947 (89.47 %)
False positive rate	0.0063 (0.63 %)	0.0066 (0.66 %)
False discovery rate	0.1030 (10. 3 %)	0.1052 (10.52%)
false omission rate	0.0357 (3.57 %)	0.0356 (3.56 %)
Positive likelihood ration	94.6186 (9461.86 %)	91.1316 (9113.16 %)
Negative likelihood ratio	27.7706 (2777.06 %)	27.8665 (2786.65 %)
prevalence threshold	0.0932 (9.32%)	0.0948 (9.48 %)
critical success index	0.5606 (56.06 %)	0.5659 (56.59 %)
F1 Score	0.7185 (71.85 %)	0.7228 (72.28 %)
Mathews Correlation co-efficient	0.7145 (71.45 %)	0.7177 (71.77 %)
Fowlkes-Mallows Index	0.7331 (73.31 %)	0.7365 (73.65 %)
informedness	0.5929 (59.29 %)	0.5996 (59.96 %)
Diagnostic odds ratio	234.6306 (23463.06 %)	229.9519 (22995.19 %)

During the testing stage, Table XII showcases the SVM model's diabetes predictions. The test dataset comprises 25,000 samples, split into 2,091 true positive cases (indicating the presence of diabetes) and 22,909 true negative cases (indicating the absence of diabetes). Among the true positive cases, the SVM model accurately identifies 1,266 samples as positive, correctly indicating the absence of healthcare issues. However, the model misclassifies 825 records as negative, erroneously suggesting the presence of healthcare issues. Conversely, among the true negative cases, the SVM model precisely predicts 22,758 samples as negative, correctly recognizing the presence of healthcare conditions. However, the model makes 151 errors, incorrectly classifying them as positive, falsely indicating the absence of healthcare issues.

Table XIII presents a comprehensive overview of the performance of the proposed SVM model in terms of various evaluation metrics. During the training phase, the SVM model achieved the following percentages for each metric: 95.98%, 4.01%, 60.47%, 99.30%, 88.98%, 0.69%, 11.01%, 3.58%, 8642.10%, 2769.42%, 9.71%, 56.26%, 72.01%, 71.44%, 73.35%, 59.77%, 21713.23%, accuracy, miss-classification rate, sensitivity, specificity, precision, False positive rate, False discovery rate, False omission rate, Positive likelihood ratio, Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio, respectively. During the validation phase, the performance of the model is evaluated, and the following evaluation metrics are obtained: 96.09%, 3.90%, 60.54%, 99.34%, 89.34%, 0.65%, 10.65%, 3.49%, 9185.62%, 2839.70%, 9.44%, 56.46%, 72.17%, 71.70%, 73.54%, 59.88%, 23127.93%, accuracy, miss-classification rate, sensitivity, specificity, precision, False positive rate, False discovery rate, False omission rate, Positive likelihood ratio, Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio, respectively.

TABLE XI. SVM MODEL'S - EHRS DIABETES DATASET – TRAINING PHASE – 75:25

Input	**Total number of samples (75000)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	6409(positive)	3876(TP)	2533(FN)
	68591(negative)	480(FP)	68111(TN)

TABLE XII. SVM MODEL'S - EHRS DIABETES DATASET – TESTING PHASE – 75:25

Input	**Total number of samples (25000)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	2091(positive)	1266(TP)	825(FN)
	22909(negative)	151(FP)	22758(TN)

TABLE XIII. SVM MODEL'S (EHRS DIABETES DATASET) EVALUATION METRICS, 75:25

	Testing	**Training**
Accuracy	0.9609 (96.09 %)	0.9598 (95.98 %)
Miss-classification rate	0.0390 (3.90 %)	0.0401 (4.01 %)
Sensitivity	0.6054 (60.54 %)	0.6047 (60.47 %)
Specificity	0.9934 (99.34 %)	0.9930 (99.30 %)
Precision	0.8934 (89.34 %)	0.8898 (88.98 %)
False positive rate	0.0065 (0.65 %)	0.0069 (0.69 %)
False discovery rate	0.1065 (10. 65 %)	0.1101 (11.01%)
false omission rate	0.0349 (3.49%)	0.0358 (3.58 %)
Positive likelihood ration	91.8562 (9185.62 %)	86.4210 (8642.10 %)
Negative likelihood ratio	28.3970 (2839.70 %)	27.6942 (2769.42 %)
prevalence threshold	0.0944 (9.44 %)	0.0971 (9.71 %)
critical success index	0.5646 (56.46 %)	0.5626 (56.26 %)
F1 Score	0.7217 (72.17 %)	0.7201 (72.01 %)
Mathews Correlation co-efficient	0.7170 (71.70 %)	0.7144 (71.44 %)
Fowlkes-Mallows Index	0.7354 (73.54 %)	0.7335 (73.35 %)
informedness	0.5988 (59.88 %)	0.5977 (59.77 %)
Diagnostic odds ratio	231.2793 (23127.93 %)	217.1323 (21713.23 %)

E. DPEMDFML - ANN System Model - using Pima Diabetes Dataset - 70:30

Shifting our focus to the second algorithm used in this research, the Artificial Neural Network (ANN) model was employed, and the Pima Diabetes Dataset was utilized for evaluation. To ensure a robust assessment of the model's effectiveness, the dataset was split into two sets: 20% for testing (n=231) and 70% for training (n=537). To gauge the performance of the ANN model, a detailed analysis was conducted using two distinct evaluation tables: Table XIV and Table XV. These tables present the confusion matrix, providing valuable insights into the model's ability to deliver accurate predictions during both the testing and training phases.

During the training stage, Table XIV illustrates the ANN model’s predictions for diabetes disease. The training dataset consists of 537 samples, further divided into 188 true positive cases, indicating the presence of diabetes, and 349 true negative cases, indicating the absence of diabetes. Among the true positive cases, the ANN model correctly identifies 157 samples as positive, accurately indicating the absence of healthcare issues. However, the model misclassifies 31 records as negative, falsely indicating the presence of healthcare issues. Conversely, among the true negative cases, the ANN model accurately predicts 327 samples as negative, correctly identifying the presence of healthcare conditions. However, the model makes 22 errors, incorrectly classifying them as positive, falsely indicating the absence of a healthcare issue.

During the testing phase, the ANN model's predictions for diabetes disease are shown in Table XV. The testing dataset consists of 231 samples, further divided into 80 true positive cases, indicating the presence of diabetes, and 151 true negative cases, indicating the absence of diabetes. Among the true positive cases, the ANN model correctly identifies 47 samples as positive, accurately indicating the absence of healthcare issues. However, the model misclassifies 33 records as negative, falsely signalling the presence of healthcare issues where there are none. On the other hand, among the true negative cases, the ANN model accurately predicts 116 samples as negative, correctly identifying the presence of healthcare conditions. However, the model makes 35 errors, incorrectly classifying them as positive, falsely indicating the absence of a healthcare issue.

Table XVI provides a comprehensive summary of the proposed ANN model's performance during the training phase, showcasing various evaluation metrics. The percentages for each metric achieved by the ANN model are as follows: 90.13% accuracy, 9.86% miss-classification rate, 83.51% sensitivity, 93.69% specificity, 87.70% precision, 6.30% false positive rate, 12.29% false discovery rate, 6.30% false omission rate, 1324.78% positive likelihood ratio, 17.59% negative likelihood ratio, 44.90% prevalence threshold, 77.20% critical success index, 85.55% F1 Score, 78.12% Mathews Correlation coefficient, 83.78% Fowlkes-Mallows Index, 77.20% informedness, and 7527.71% diagnostic odds ratio. During the validation phase, the performance of the model is evaluated, and the following evaluation metrics are obtained: 70.56%, 29.43%, 58.75%, 76.82%, 57.31%, 23.17%, 42.68%, 23.17%, 253.46%, 53.69%, 40.96%, 35.57%, 58.02%, 35.36%, 61.55%, 35.57%, 472.03%, accuracy, miss-classification rate, sensitivity, specificity, precision, False positive rate, False discovery rate, False omission rate, Positive likelihood ratio, Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio, respectively.

TABLE XIV. ANN MODEL'S - PIMA DIABETES DATASET – TRAINING PHASE – 70:30

Input	***Total number of samples (537)***	***Result (output)***	
	Expected output	*Predicted positive*	*Predicted negative*
	188(positive)	*157(TP)*	*31(FN)*
	349 (negative)	*22(FP)*	*327 (TN)*

TABLE XV. ANN MODEL'S - PIMA DIABETES DATASET – TESTING PHASE – 70:30

Input	**Total number of samples (231)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	80(positive)	47(TP)	33(FN)
	151(negative)	35(FP)	116(TN)

TABLE XVI. ANN MODEL'S (PIMA DIABETES DATASET) EVALUATION METRICS, 70:30

	Testing	**Training**
Accuracy	0.7056 (70.56 %)	0.9013 (90.13%)
Miss-classification rate	0.2943 (29.43 %)	0.0986 (9.86 %)
Sensitivity	0.5875 (58.75 %)	0.8351 (83.51%)
Specificity	0.7682 (76.82 %)	0.9369 (93.69 %)
Precision	0.5731 (57.31 %)	0.8770 (87.70%)
False positive rate	0.2317 (23.17 %)	0.0630 (6.30 %)
False discovery rate	0.4268 (42. 68 %)	0.1229 (12.29 %)
false omission rate	0.2317 (23.17%)	0.06303 (6.30 %)
Positive likelihood ration	2.5346 (253.46 %)	13.2478 (1324.78 %)
Negative likelihood ratio	0.5369 (53.69 %)	0.1759 (17.59 %)
prevalence threshold	0.4096 (40.96 %)	0.4490 (44.90 %)
critical success index	0.3557 (35.57 %)	0.7720 (77.20 %)
F1 Score	0.5802 (58.02 %)	0.8555 (85.55 %)
Mathews Correlation co-efficient	0.3536 (35.36 %)	0.7812 (78.12 %)
Fowlkes-Mallows Index	0.6155 (61.55 %)	0.8378 (83.78 %)
Informedness	0.3557 (35.57 %)	0.7720 (77.20 %)
Diagnostic odds ratio	4.7203 (472.03 %)	75.2771 (7527.71 %)

F. DPEMDFML - ANN System Model - using Pima Diabetes Dataset - 75:25

Once more, the ANN model was utilized with the Pima Diabetes Dataset. The dataset here was split into 25% for testing (n=192) and 75% for training (n=576) to ensure a thorough evaluation of the model's performance. The performance metrics of the ANN model are presented in Table XVII and Table XVIII, displaying the confusion matrix results.

During the training phase, Table XVII showcases the ANN model's predictions for diabetes disease. Out of the 576 samples used for training, 199 are identified as real positive cases, and 377 as real negative cases. Among these, 172 are correctly identified as positive, meaning no healthcare issues have been observed, while 27 are incorrectly projected as negatives, indicating a healthcare issue is present. Regarding the 377 samples with negative results, indicating the presence of a healthcare condition, 352 samples are correctly forecasted as negative, and 25 samples are wrongly forecasted as positive, indicating the absence of a healthcare issue.

During the testing phase, Table XVIII displays the ANN model's predictions for diabetes disease. The dataset consists of 192 samples, divided into 69 real positive cases and 123 real negative cases. Among these, the model correctly identifies 45 samples as positive, indicating no healthcare issues observed, while 24 samples are incorrectly projected as negatives, suggesting a healthcare issue. For the 123 samples with negative results, indicating the presence of a healthcare condition, the model appropriately forecasts 92 as negative, and 31 samples are wrongly forecasted as positive, indicating the absence of a healthcare issue.

Table XIX provides a comprehensive summary of the proposed ANN model's performance during the training phase, showcasing various evaluation metrics. The percentages for each metric achieved by the ANN model are as follows: 90.97%, 9.02%, 86.43%, 93.36%, 87.30%, 6.63%, 12.96%, 5.88%, 1303.39%, 14.53%, 46.53%, 79.80%, 86.86%, 79.99%, 84.98%, 79.80%, 8969.48%, accuracy, miss-classification rate, sensitivity, specificity, precision, False positive rate, False discovery rate, False omission rate, Positive likelihood ratio, Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio, respectively. During the validation phase, the performance of the model is evaluated, and the following evaluation metrics are obtained: 71.35%, 28.64%, 65.21%, 74.79%, 59.21%, 25.20%, 40.78%, 20.68%, 258.76%, 46.50%, 45.21%, 40.01%, 62.06%, 39.26%, 61.10%, 40.01%, 556.45%, accuracy, miss-classification rate, sensitivity, specificity, precision, False positive rate, False discovery rate, False omission rate, Positive likelihood ratio, Negative likelihood ratio, Prevalence threshold, critical success index, F1 Score, Mathews Correlation coefficient, Fowlkes-Mallows Index, informedness, and Diagnostic odds ratio, respectively.

G. DPEMDFML - ANN System Model - using EHRs Dataset - 70:30

Utilizing the same algorithm, the ANN model applied to the second dataset, referred to as the EHRs Dataset (Electronic Health Records Dataset). To achieve a comprehensive evaluation of the model's performance, the data set was split as: 30% for testing (n = 30,000) and 70% for training (n = 70,000). The effectiveness of the ANN model was assessed through a thorough analysis of its performance using two separate evaluation tables: Table XX and Table XXI. These tables present detailed information from the confusion matrix, offering insights into the model's performance during both the testing and training phases.

During the training phase, Table XX displays the outcomes of the ANN model's predictions for diabetes disease. In this phase, the model uses a dataset consisting of 70,000 samples, which are further divided into 5,972 real positive cases and 64,028 real negative cases. Among the real positive cases, 4,265 samples are correctly identified as positive, indicating the absence of healthcare issues. However, 1,707 samples are incorrectly classified as negatives, implying potential healthcare concerns. Regarding the real negative cases, which represent the presence of a healthcare condition, the model accurately predicts 63,938 samples as negative, indicating the presence of healthcare issues. However, 90 samples are falsely predicted as positive, suggesting the absence of healthcare issues, when in fact, they should have been classified as negative.

TABLE XVII. ANN MODEL'S - PIMA DIABETES DATASET – TRAINING PHASE – 75:25

Input	**Total number of samples (576)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	199(positive)	172(TP)	27(FN)
	377 (negative)	25(FP)	352(TN)

TABLE XVIII. ANN MODEL'S - PIMA DIABETES DATASET – TESTING PHASE – 75:25

Input	**Total number of samples (192)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	69(positive)	45(TP)	24(FN)
	123(negative)	31(FP)	92(TN)

TABLE XIX. ANN MODEL'S (PIMA DIABETES DATASET) EVALUATION METRICS, 75:25

	Testing	Training
Accuracy	0.7135 (71.35 %)	0.9097 (90.97%)
Miss-classification rate	0.2864 (28.64 %)	0.0902 (9.02 %)
Sensitivity	0.6521 (65.21 %)	0.8643 (86.43%)
Specificity	0.7479 (74.79 %)	0.9336 (93.36 %)
Precision	0.5921 (59.21 %)	0.8730 (87.30%)
False positive rate	0.2520 (25.20 %)	0.0663 (6.63 %)
False discovery rate	0.4078 (40.78 %)	0.1269 (12.96 %)
false omission rate	0.2068 (20.68 %)	0.0588 (5.88 %)
Positive likelihood ration	2.5876 (258.76 %)	13.0339 (1303.39 %)
Negative likelihood ratio	0.4650 (46.50 %)	0.1453 (14.53 %)
prevalence threshold	0.4521 (45.21 %)	0.4653 (46.53 %)
critical success index	0.4001 (40.01 %)	0.7980 (79.80 %)
F1 Score	0.6206 (62.06 %)	0.8686 (86.86 %)
Mathews Correlation co-efficient	0.3926 (39.26 %)	0.7999 (79.99 %)
Fowlkes-Mallows Index	0.6110 (61.10 %)	0.8498 (84.98 %)
informedness	0.4001 (40.01 %)	0.7980 (79.80 %)
Diagnostic odds ratio	5.5645 (556.45 %)	89.6948 (8969.48 %)

During the testing phase, Table XXI demonstrates the ANN model's performance in predicting diabetes disease. The dataset used for testing consists of 30,000 samples, which are further divided into 2,528 actual positive cases and 27,472 actual negative cases. The model correctly identifies 1,754 positive cases, indicating the absence of healthcare issues. However, it mistakenly classifies 774 positive cases as negative, suggesting possible healthcare concerns. For the actual negative cases, which indicate the presence of healthcare conditions, the model accurately predicts 27,368 samples as negative. This demonstrates its ability to identify the presence of healthcare issues correctly. Nevertheless, there are 104 false positive predictions, where the model incorrectly identifies cases as negative, indicating the absence of healthcare issues when they should have been classified as positive.

Table XXII provides a comprehensive summary of the proposed ANN model's performance during the training phase, showcasing various evaluation metrics. The percentages for each metric achieved by the ANN model are as follows: 97.43% accuracy, 2.56% miss-classification rate, 71.41% sensitivity, 99.85% specificity, 97.93% precision, 0.14% false positive rate, 2.06% false discovery rate, 2.60% false omission rate, 50807.36% positive likelihood ratio, 28.62% negative likelihood ratio, 35.77% prevalence threshold, 71.27% critical success index, 82.59% F1 Score, 82.43% Mathews Correlation coefficient, 97.12% Fowlkes-Mallows Index, and 177501.51% diagnostic odds ratio. During the validation phase, the performance of the model is evaluated, and the following evaluation metrics are obtained: 97.07% accuracy, 2.92% miss-classification rate, 69.38% sensitivity, 99.62% specificity, 94.40% precision, 0.37% false positive rate, 5.59% false discovery rate, 2.75 % false omission rate, 18327.76% positive likelihood ratio, 30.733% negative likelihood ratio, 34.88% prevalence threshold, 69.00% critical success index, 79.98% F1 Score, 79.52% Mathews Correlation coefficient, 96.73% Fowlkes-Mallows Index, and 59634.60% diagnostic odds ratio.

TABLE XX. ANN MODEL'S - EHRS DIABETES DATASET – TRAINING PHASE – 70:30

Input	**Total number of samples (70000)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	5972(positive)	4265 (TP)	1707 (FN)
	64028 (negative)	90 (FP)	63938(TN)

TABLE XXI. ANN MODEL'S - EHRS DIABETES DATASET – TESTING PHASE – 70:30

Input	**Total number of samples (30000)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	2528(positive)	1754 (TP)	774 (FN)
	27472(negative)	104 (FP)	27368 (TN)

TABLE XXII. ANN MODEL'S (EHRS DIABETES DATASET) EVALUATION METRICS, 70:30

	Testing	**Training**
Accuracy	0.9707 (97.07 %)	0.9743 (97.43 %)
Miss-classification rate	0.0292 (2.92 %)	0.0256 (2.56 %)
Sensitivity	0.6938 (69.38 %)	0.7141 (71.41 %)
Specificity	0.9962 (99.62 %)	0.9985 (99.85 %)
Precision	0.9440 (94.40 %)	0.9793 (97.93 %)
False positive rate	0.0037 (0.37 %)	0.0014 (0.14 %)
False discovery rate	0.0559 (5.59 %)	0.0206 (2.06 %)
false omission rate	0.0275 (2.75 %)	0.0260 (2.60 %)
Positive likelihood ration	183.2776 (18327.76 %)	508.0736 (50807.36 %)
Negative likelihood ratio	0.30733 (30.733 %)	0.2862 (28.62 %)
prevalence threshold	0.3488 (34.88 %)	0.3577 (35.77 %)
critical success index	0.6900 (69.00 %)	0.7127 (71.27 %)
F1 Score	0.7998 (79.98 %)	0.8259 (82.59 %)
Mathews Correlation co-efficient	0.7952 (79.52 %)	0.8243 (82.43 %)
Fowlkes-Mallows Index	0.9673 (96.73 %)	0.9712 97.12 %)
Informedness	0.6900 (69.00 %)	0.7127 (71.27 %)
Diagnostic odds ratio	596.3460 (59634.60 %)	1775.0151 (177501.51 %)

H. *DPEMDFML - ANN System Model - using EHRs Dataset - 75:25*

In this study, the Artificial Neural Network (ANN) model was utilized to analyse the Electronic Health Records Dataset (EHRs Dataset). To ensure a rigorous evaluation of the model's capabilities, the dataset was split into 25% for testing, comprising 25,000 samples, and 75% for training, with 75,000 samples. The effectiveness of the ANN model was thoroughly assessed using two distinct evaluation tables: Table XXIII and Table XXIV, which offer a detailed view of the confusion matrix and facilitate an in-depth analysis of the model's performance.

During the training phase, Table XXIII depicts the predictions made by the ANN model for diabetes disease. The dataset used for training consists of 75,000 samples, which are further categorized into 6,409 real positive cases and 68,591 real negative cases. The model accurately identified 4,582 samples as truly positive, indicating the absence of healthcare issues. However, it misclassified 1,827 records as negatives, falsely signalling the presence of a healthcare condition. Out of the 68,591 negative results, which indicate the presence of a healthcare condition, the model correctly forecasted 68,472 samples as negative, demonstrating its effectiveness in correctly identifying such cases. However, there were 119 samples that were inaccurately forecasted as positive, indicating the absence of a healthcare issue when it was present.

During the testing phase, Table XXIV presents the predictions made by the ANN model for diabetes disease. The dataset used for testing comprises 25,000 samples, which are further divided into 2,091 real positive cases and 22,909 real negative cases. The model accurately identified 1,461 samples as truly positive, indicating the absence of healthcare issues. However, it misclassified 630 records as negatives, falsely signaling the presence of a healthcare condition. Out of the 22,909 negative results, which indicate the presence of a healthcare condition, the model correctly forecasted 22,827 samples as negative, demonstrating its effectiveness in correctly identifying such cases. However, there were 82 samples that were inaccurately forecasted as positive, indicating the absence of a healthcare issue when it was present.

Table XXV provides a comprehensive summary of the ANN model's performance during the training phase, displaying various evaluation metrics. The ANN model achieved the following percentages for each metric: 97.40% for accuracy, 2.59% for miss-classification rate, 71.49% for sensitivity, 99.82% for specificity, 97.96% for precision, 0.17% for the False positive rate, 2.53% for the False discovery rate, 28.50% for the False omission rate, 41208.32% for the Positive likelihood ratio, 28.55% for the Negative likelihood ratio, 35.83% for the Prevalence threshold, 71.31% for the critical success index, 82.48% for the F1 Score, 82.25% for the Mathews Correlation coefficient, 97.09% for the Fowlkes-Mallows Index, 71.31% for informedness, and 144305.40% for the Diagnostic odds ratio. During the testing phase, the ANN model achieved the following percentages for each evaluation metric: 97.51% for accuracy, 2.84% for miss-classification rate, 69.87% for sensitivity, 99.64% for specificity, 94.68% for precision, 0.35% for the False positive rate, 5.51% for the False discovery rate, 30.12% for the False omission rate, 19520.38% for the Positive likelihood ratio, 30.23% for the Negative likelihood ratio, 35.11% for the Prevalence threshold, 69.51% for the critical success index, 80.40% for the F1 Score, 79.96% for the Mathews Correlation coefficient, 96.82% for the Fowlkes-Mallows Index, 69.51% for informedness, and 64557.19% for the Diagnostic odds ratio.

TABLE XXIII. ANN MODEL'S - EHRS DIABETES DATASET – TRAINING PHASE – 75:25

Input	**Total number of samples (75000)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	6409(positive)	4582 (TP)	1827 (FN)
	68591 (negative)	119 (FP)	68472(TN)

TABLE XXIV. ANN MODEL'S - EHRS DIABETES DATASET – TESTING PHASE – 75:25

Input	**Total number of samples (25000)**	**Result (output)**	
	Expected output	Predicted positive	Predicted negative
	2091(positive)	1461 (TP)	630 (FN)
	22909(negative)	82 (FP)	22827 (TN)

TABLE XXV. ANN MODEL'S (EHRS DIABETES DATASET) EVALUATION METRICS, 75:25

	Testing	**Training**
Accuracy	0.9715 (97.51 %)	0.9740 (97.40 %)
Miss-classification rate	0.0284 (2.84 %)	0.0259 (2.59 %)
Sensitivity	0.6987 (69.87 %)	0.7149 (71.49 %)
Specificity	0.9964 (99.64 %)	0.9982 (99.82 %)
Precision	0.9468 (94.68 %)	0.9746 (97.96 %)
False positive rate	0.0035 (0.35 %)	0.0017 (0.17 %)
False discovery rate	0.0531 (5.51 %)	0.0253 (2.53 %)
false omission rate	0.3012 (30.12 %)	0.2850 (28.50 %)
Positive likelihood ration	195.2038 (19520.38 %)	412.0832 (41208.32 %)
Negative likelihood ratio	0.3023 (30.23 %)	0.2855 (28.55 %)
prevalence threshold	0.3511 (35.11 %)	0.3583 (35.83 %)
critical success index	0.6951 (69.51 %)	0.7131 (71.31 %)
F1 Score	0.8040 (80.40 %)	0.8248 (82.48 %)
Mathews Correlation co-efficient	0.7996 (79.96 %)	0.8225 (82.25 %)
Fowlkes-Mallows Index	0.9682 (96.82 %)	0.9709 97.09 %)
Informedness	0.6951 (69.51 %)	0.7131 (71.31 %)
Diagnostic odds ratio	645.5719 (64557.19 %)	1443.0540 (144305.40 %)

The results of DPEMDFML model on the EHRs diabetes dataset indicate that the ANN model outperformed other algorithms in both the 70:30 and 75:25 ratio splits. With the 70:30 split, the ANN model achieved an impressive accuracy of 97.43%, showcasing its robustness in correctly classifying diabetes cases.

Similarly, in the 75:25 split, the ANN model maintained a high accuracy of 97.40%, further validating its effectiveness in handling the dataset. On the other hand, the SVM model also showcased commendable results on the same EHRs diabetes dataset. In the 70:30 split, the SVM model achieved an accuracy of 96.03%, demonstrating its potential to effectively classify diabetes cases.

In the 75:25 split, the SVM model maintained a high accuracy of 95.98%, further highlighting its capability to handle varying data proportions. Table XXVI show the accuracies reached in this study.

TABLE XXVI. PERFORMANCE OF PROPOSED DPEMDFML MODEL W.R.T PIMA DATASET AND EHRS DATASET

	PIMA dataset 70:30	**EHRs dataset 70:30**	**PIMA dataset 75:25**	**EHRs dataset 75:25**
SVM	74.46 %	96.03%	78.81%	95.98%
ANN	90.13%	97.43%	90.97%,	97.40%

Table XXVII presented provides an overall comparison of the proposed DPEMDFML model with the previous works mentioned. The results clearly demonstrate that the accuracy of the proposed model has outperformed all the other accuracies reported in the mentioned works, using both of the employed algorithms.

TABLE XXVII. COMPARISON OF PROPOSED DPEMDFML MODEL WITH PREVIOUS WORKS MENTIONED

Research Study	**Method**	**Accuracy**
Akkarapol and Jongsawas [11]		77.11%
Kavakiotis et al. [12]		84%
Xue-Hui Meng et al. [13]	• Logistic Regression Model • Decision Tree Model (C5.0) • Artificial Neural Networks (ANN) Model	76.13% 77.87% 73.23%
Dey et al. [15]	• ANN Model with MMS	82.35%
Proposed DPEMDFML model	• ANN • SVM	97.43% 96.03%

VI. CONCLUSION

In summary, this research offers a distinctive and thorough investigation of the application of machine learning approaches for diabetes detection. The proposed DPEMDFML model shows improved accuracy in predicting diabetes disease compared to earlier efforts by using two separate algorithms and two different datasets. The comprehensive assessment tables show that the SVM and ANN models performed well during both the testing and training periods. The suggested framework's use of machine learning fusion has the potential to diagnose diabetes earlier, resulting in proactive healthcare treatments and better patient outcomes. This work advances the field of diabetes diagnostic research by offering insightful information on the efficacy of various algorithms and datasets. The findings open the way for further study and model enhancement, with the goal of facilitating improved and more accurate diabetes detection in clinical situations. In future, we will incorporate more recent datasets to enhance the study's relevance and accuracy.

REFERENCES

[1] [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1.

[2] Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B. J. & Lernmark, A. Type 1 diabetes mellitus. Nature reviews Disease primers, vol. 3,no. 1, pp. 1-17, 2017.

[3] https://www.cdc.gov/diabetes/basics/prediabetes.html

[4] F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, Computer Vision and Machine Intelligence in Medical Image Analysis. London, U.K.: Springer, 2019

[5] Rehman, A., Athar, A., Khan, M. A., Abbas, S., Fatima, A., & Saeed, A. Modelling, simulation, and optimization of diabetes type II prediction using deep extreme learning machine. Journal of Ambient Intelligence and Smart Environments,vol. 12, no. 2, pp. 125-138, 2020.

[6] Joyia, G. J., Liaqat, R. M., Farooq, A., & Rehman, S. Internet of medical things (IoMT): Applications, benefits and future challenges in healthcare domain. J. Commun., vol. 12, no. 4, pp. 240-247, 2017.

[7] Muneer S, Rasool MA. A Enhancing Healthcare Outcomes with Explainable AI (XAI) for Disease Prediction: A Comprehensive Review. International Journal of Advanced Sciences and Computing, vol. 1, no. 1, pp. 37-42, 2022.

[8] Siddiqui, S. Y., Haider, A., Ghazal, T. M., Khan, M. A., Naseer, I., Abbas, S. & Ateeq, K. IoMT cloud-based intelligent prediction of breast cancer stages empowered with deep learning. IEEE Access, vol. 9, pp. 146478-146491, 2021.

[9] Jakkula, V. Tutorial on support vector machine (svm). School of EECS, Washington State University, vol. 37 no. 2, pp. 3-7, 2006.

[10] Boser, B. E., Guyon, I. M., & Vapnik, V. N.. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152, 1992.

[11] Sa-ngasoongsong, A., & Chongwatpol, J. An analysis of diabetes risk factors using data mining approach. Oklahoma state university, USA, pp.1-55, 2012.

[12] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, vol. 15, pp. -116, 2017.

[13] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung Journal of Medical Sciences, vol. 29, no., pp. 93-99, 2013

[14] Muneer SM, Alvi MB, Farrakh A. Cyber Security Event Detection Using Machine Learning Technique. International Journal of Computational and Innovative Sciences, vol. 2, no. 2, pp. 42-46, 2023.

[15] Dey, S. K., Hossain, A., & Rahman, M. M. Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm. In 2018 21st international conference of computer and information technology (ICCIT) IEEE, pp. 1-5, 2018.

[16] Pradhan, G., Pradhan, R., & Khandelwal, B. A study on various machine learning algorithms used for prediction of diabetes mellitus. In Soft Computing Techniques and Applications: Proceeding of the International Conference on Computing and Communication (IC3 2020) Springer Singapore, pp. 553-561, 2021.

[17] Abbas, T., Fatima, A., Shahzad, T., Alissa, K., Ghazal, T. M., Al-Sakhnini, M. M., & Ahmed, A. Secure IoMT for disease prediction empowered with transfer learning in healthcare 5.0, the concept and case study. IEEE Access, vol. 11, pp. 39418 – 39430, 2023.

[18] Abbas, S., Issa, G. F., Fatima, A., Abbas, T., Ghazal, T. M., Ahmad, M., & Khan, M. A. Fused Weighted Federated Deep Extreme Machine Learning Based on Intelligent Lung Cancer Disease Prediction Model for Healthcare 5.0. International Journal of Intelligent Systems, vol. 2023, pp. 1-15, 2023.

[19] Asif, R. N., Abbas, S., Khan, M. A., Sultan, K., Mahmud, M., & Mosavi, A. Development and validation of embedded device for electrocardiogram arrhythmia empowered with transfer learning. Computational Intelligence and Neuroscience, vol. 2022, pp. 1-14, 2022.

[20] Arooj, S., Zubair, M., Khan, M. F., Alissa, K., Khan, M. A., & Mosavi, A. Breast cancer detection and classification empowered with transfer learning. Frontiers in Public Health, vol. 10, pp. 1- 19, 2022.

[21] Khan, M. B. S., Nawaz, M. S., Ahmed, R., Khan, M. A., & Mosavi, A. Intelligent breast cancer diagnostic system empowered by deep extreme gradient descent optimization. Mathematical Biosciences and Engineering, vol. 19, no. 8, pp. 7978-8002, 2022.

[22] Ahmad, M., Alfayad, M., Aftab, S., Khan, M. A., Fatima, A., Shoaib, B., & Elmitwal, N. S. Data and Machine Learning Fusion Architecture for Cardiovascular Disease Prediction. Computers, Materials & Continua, vol. 69, no. 2, pp. 2717-2730, 2021.

[23] Siddiqui, S. Y., Naseer, I., Khan, M. A., Mushtaq, M. F., Naqvi, R. A., Hussain, D., & Haider, A. Intelligent breast cancer prediction empowered with fusion and deep learning. Computers, Materials and Continua, vol. 67, no. 1, pp. 1033-1049, 2021.