A Home-Based Diabetes Prediction System on Internet of Things, Federated Learning and Edge Computing

Long Huynh-Phi[1,2], Duy Nguyen-Khanh[1,2], Thuat Nguyen-Khanh[1,2(✉)], Chuong Dang-Le-Bao[1,2], and Quan Le-Trung[1,2]

[1] Faculty of Computer Networks and Communications, University of Information Technology, Ho Chi Minh City, Viet Nam
{20521562,20521240}@gm.uit.edu.vn, {thuatnk,chuongdlb,quanlt}@uit.edu.vn
[2] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. Detecting the disease early is an important step in reducing its impact. In recent years, applications that monitor and predict health metrics using machine learning have attracted public attention. Our research built a diabetes health monitoring and prediction system based on the Edge Computing model. For hospitals, users and patients are represented by K3 clusters. The K-Nearest Neighbor (KNN) algorithm is run in a distributed fashion using Federated Learning with the proposed system. It could allow people to track their vital health indicators without having to go to the hospital. In our proposed system, the diabetes risk level can be predicted in advance so that users can take preventive steps. Through Federated Learning used, the model is being trained on distributed data sources guaranteed to preserve privacy and improve accuracy. K-Nearest Neighbor in the federated learning cluster, we can improve the prediction accuracy by up to 10% compared to the standalone version of K-Nearest Neighbor.

Keywords: Home-based health system · Internet of Things · Federated learning · Edge computing · Diabetes prediction

1 Introduction

Diabetes, a chronic medical condition affecting millions worldwide, necessitates innovative solutions for effective management. A Home-based Diabetes Prediction system has emerged as a groundbreaking technological advancement. Utilizing cutting-edge technologies like the Internet of Things (IoT), Federated Learning, and Edge Computing, this system aims to transform diabetes management by offering predictive capabilities and healthcare services within the confines of patients' homes.

One approach to tackle the challenge of diabetes management is through predictive analytic powered by KNN algorithms. By integrating data from diverse sources such as wearable devices, glucose monitors, and health applications. The

N.-S. Vo et al. (Eds.): INISCOM 2024, LNICST 595, pp. 14–27, 2024.
https://doi.org/10.1007/978-3-031-67357-3_2

system can analyze this information to predict the risk of diabetes. Model KNN, trained on this extensive and a variety of datasets, can identify patterns and trends, enabling early detection and prevention.

Federated Learning ensures privacy by training models across decentralized devices, keeping patient data local. Through collaborative learning, it preserves individual privacy, enhancing data security and user confidence. This approach enables personalized care plans, analyzing historical and real-time data for customized recommendations. Remote monitoring empowers healthcare professionals to intervene promptly, enhancing care quality and encouraging patient involvement in their health management journey.

The current body of research has extensively explored remote health monitoring, diabetes prediction using KNN, and federated learning in healthcare as standalone areas. However, a notable deficiency exists in the integration of these domains. Our study bridges this gap by developing a unified remote health monitoring platform that employs KNN for diabetes prediction. Through the incorporation of federated learning, we not only augment predictive accuracy but also safeguard user data privacy. This amalgamation of techniques in our system not only furnishes precise health insights but also delivers personalized predictions, effectively mitigating privacy concerns associated with sensitive health data.

Our main contributions are summarized as follows:

- Design and deploy an health monitoring system on Libelium hardware;
- Develop federated learning for prediction diabetes function;
- Build the Kubernetes cluster to save users data as so as federated learning models;
- Develop the user interfaces for two user groups: patients and medical facilities.

The rest of this paper is constructed as follows: Sect. 2 highlights the previous works on IoT-based and machine learning-based monitoring systems; The gap in e-healthcare and our motivation are also described at this section; Our proposal is showed at Sect. 3, it includes system architecture and work flows; The test-cases and performance evaluation are introduced at Sect. 4; The paper ends with our conclusions and future works.

2 Related Works

2.1 Remote Health Monitoring Systems

With the robust development of IoT and artificial intelligence has paved the way for transformative advancements across various domains, including healthcare. As the population increases rapidly, smart cities are recognizing the necessity of implementing home health monitoring systems [1]. The effectiveness of such systems is particularly evident during the COVID-19 pandemic when widespread lock downs were imposed [2]. Alongside the rapid progress in medical sensor technology [3], the establishment of home health monitoring systems has become even more convenient, emphasizing their importance in modern smart cities.

2.2 Diabetes Prediction Using Machine Learning

Machine Learning (ML) is becoming increasingly popular in the field of healthcare and is expected to be an effective tool in preventive at-home healthcare [4]. In the medical field, the application of machine learning techniques to predict diabetes based on health data has shown significant potential. Particularly, using model KNN to detect diabetic retinopathy from retinal images is an emerging and promising area [5]. The K-Nearest Neighbors (KNN) model is one of the essential tools in machine learning with numerous advantages such as low computational cost, high durability, generalization capabilities, high performance, simplicity, and ease of understanding. Employing KNN for classifying retinal images can assist in diabetes diagnosis and monitoring disease progression effectively [6]. This demonstrates the considerable potential of KNN in enhancing healthcare quality and preventing diabetes.

2.3 Federated Learning for Privacy-Preserving Healthcare

Federated learning, as highlighted in the reference [8], has become a pivotal solution for mitigating privacy concerns in healthcare data sharing, a critical issue discussed in [7]. This innovative approach revolutionizes the landscape of data-driven healthcare by enabling model training across a multitude of decentralized data sources without necessitating the sharing of raw, sensitive patient information.

One notable contribution to the field comes from Li et al., detailed in their comprehensive work [9]. They introduced a sophisticated federated learning framework meticulously designed for disease prediction, drawing upon electronic health records (EHRs) [10] sourced from diverse hospitals. What sets this framework apart is its ability to empower model KNN to collaboratively learn and improve prediction accuracy without compromising individual data privacy.

3 System Design

3.1 Architecture of the Health Monitoring System

In the architecture of our health monitoring system, we have strategically adopted a modular approach, specifically the Edge Computing paradigm, which encompasses three integral components: the cloud layer, the edge layer, and the device layer. This meticulous design aims to optimize performance, enhance scalability, and provide a seamless user experience.

The Cloud Layer is the topmost tier in our health monitoring system, managing data from the edge and device layers. It serves as a centralized hub for data storage, analytics, and user interfaces. Historical health data is stored for long-term analysis, using advanced algorithms like KNN for insights and predictions. User interfaces, such as web-based dashboards and mobile apps, enable users and healthcare professionals to access information and make informed decisions.

Strong security measures protect sensitive health data during transmission and storage.

The Edge Layer acts as a mediator between the cloud and device layers, preprocessing and filtering data before sending it to the cloud. This reduces latency and enhances real-time monitoring. Raw data is cleaned, filtered, and aggregated here, minimizing the data sent to the cloud. Basic analysis and anomaly detection enable swift responses to health events. Edge devices, like gateways or servers, are strategically placed to manage data flow efficiently.

The Device Layer is crucial in our health monitoring system, housing devices like fitness trackers, medical sensors, and IoT devices. These gadgets collect real-time health data such as heart rate, blood pressure, temperature, and activity levels. The gathered data is sent securely to the edge layer for preprocessing and then to the cloud layer for storage and analysis. To ensure uninterrupted monitoring, these devices are designed with energy-efficient technologies, extending their battery life. Additionally, a variety of sensors and technologies are integrated into these devices to monitor diverse health parameters (Fig. 1).

Our health monitoring system, utilizing Edge Computing, optimizes performance and scalability with a modular approach. The Cloud Layer enables advanced analyses, the Edge Layer reduces latency for real-time monitoring, and the energy-efficient Device Layer ensures continuous data collection. This flexible and modern solution meets the dynamic needs of users and healthcare professionals in an era emphasizing personalized and active health.

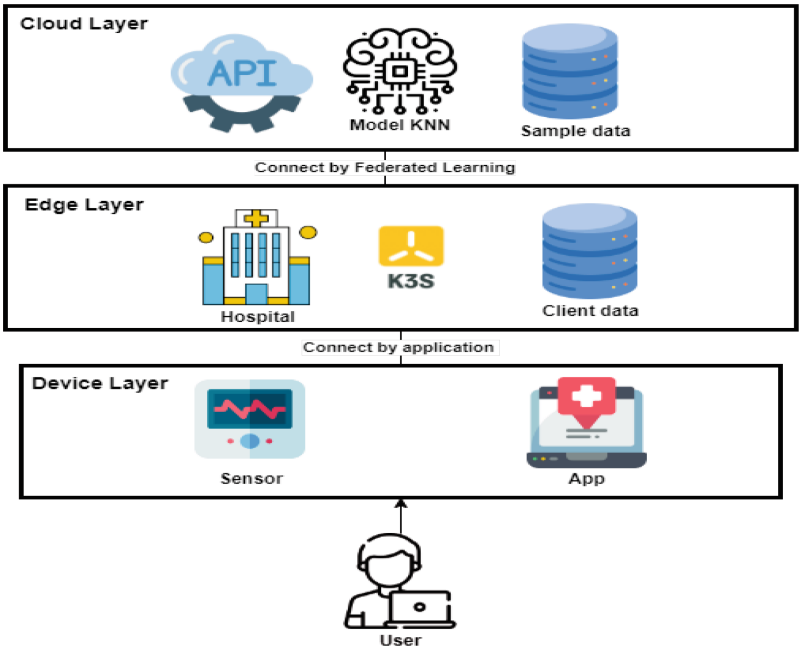

Fig. 1. Architecture of System.

3.2 User Interface and Experience

Our system's user interface prioritizes simplicity, usability, and clear data visualization. Accessible via mobile app or web portal, it displays real-time health metrics, historical trends, and personalized recommendations. Interactive visualizations and intuitive navigation enhance the user experience. Users can customize charts, graphs, and alerts and receive notifications for critical health events, like sudden glucose level changes (Fig. 2).

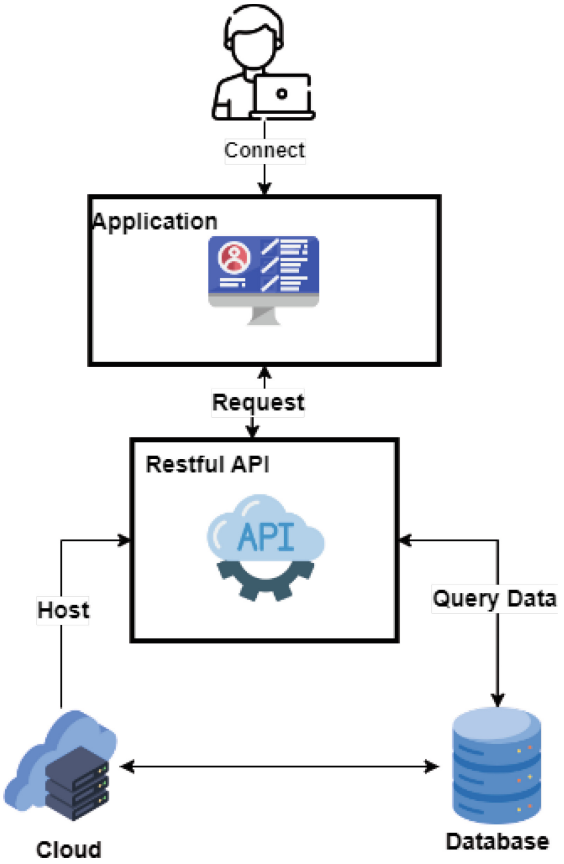

Fig. 2. Application diagram.

The user interface provides a comprehensive view of the individual's health journey. Upon logging in, users are greeted with a personalized dashboard that prominently showcases their name, creating a sense of familiarity and tailored experience. Alongside the user's name, vital health metrics are prominently displayed, offering real-time insights into their overall well-being. These metrics

may include data points such as heart rate, blood pressure, glucose levels, activity levels, and more, depending on the monitoring parameters set within the system.

In addition to the immediate health metrics, the interface offers a dedicated feature for diabetes monitoring. A prominent and easily accessible button labeled "Diabetes Diagnosis" is present, allowing users to initiate a diabetes assessment. Upon clicking this button, the system may prompt users to input specific health data relevant to diabetes diagnosis, such as fasting blood sugar levels or other pertinent information.

The user interface is designed with a focus on user-friendliness and efficiency. The diabetes diagnosis button serves as a proactive tool, empowering users to assess their diabetes risk conveniently. This intuitive design encourages regular health check-ins, fostering a proactive approach to health management. Overall, the interface not only provides essential health data at a glance but also facilitates active engagement and health-conscious decision-making through its user-friendly features (Fig. 3).

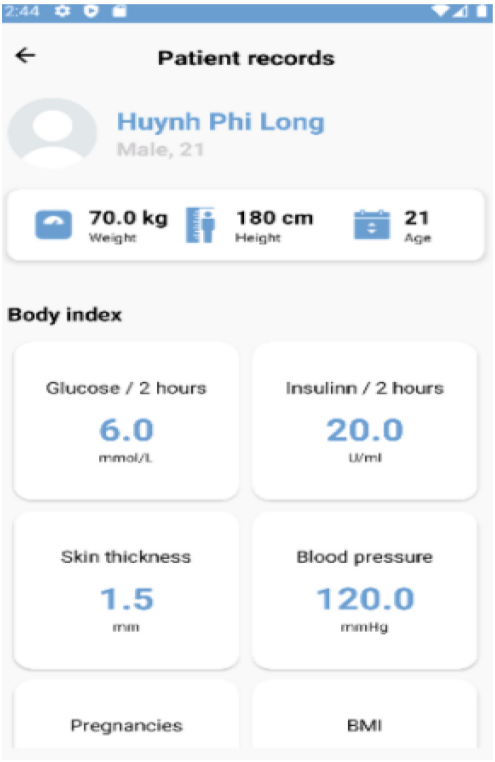

Fig. 3. User interface.

3.3 Architecture of Federated Learning

We'll start by using existing data to craft our training weights, employing the powerful K-Nearest Neighbors (KNN) algorithm and GridSearchCV. This process occurs on a cloud-based platform, ensuring scalability and computational efficiency. As we venture further into the process, the heart of our system will revolve around the implementation of Federated Learning (FL), facilitated by the seamless communication capabilities of sockets. These sockets will serve as the conduits through which data flows securely and efficiently, enabling the transfer of knowledge between our centralized cloud platform and a network of diverse healthcare institutions (Fig. 4).

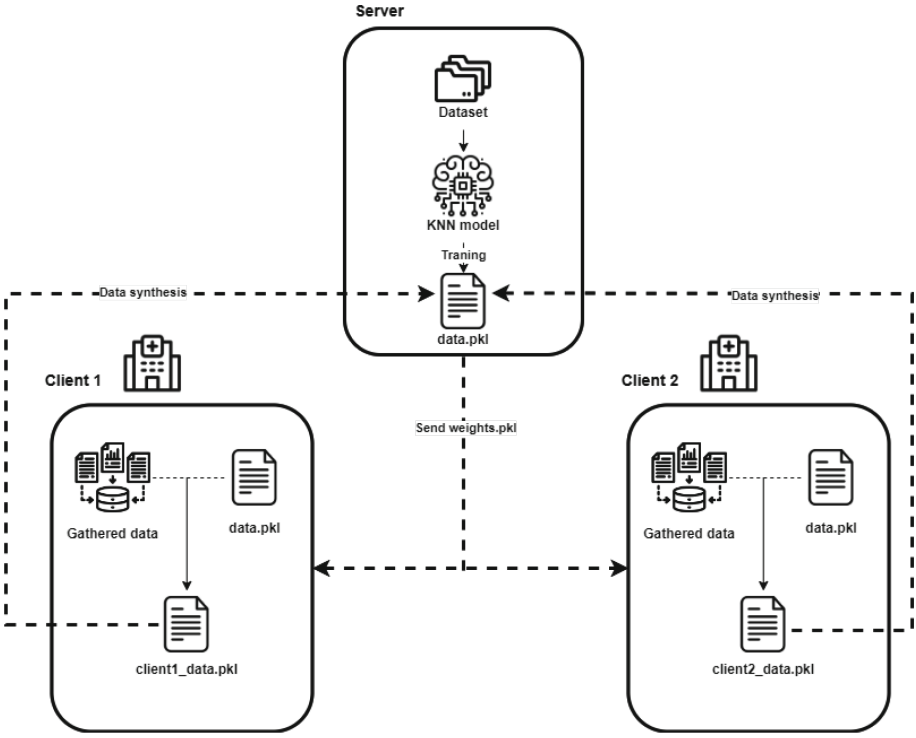

Fig. 4. Federated Learning.

Within every healthcare facility, a crucial transformation takes place, driven by cutting-edge technology and collaborative efforts. The process begins with the reception of training weights sourced from the cloud through secure socket connections. These weights play a pivotal role, guiding the local training process within the confines of each hospital.

During this localized training phase, the model undergoes refinement, honing its parameters using the institution's specific and unique dataset. This intricate

process ensures that the model becomes intimately acquainted with the intricacies of the data it is meant to analyze. Through continuous iterations, the model updates its training weights, enhancing its accuracy and relevance.

After local training, refined weights with insights from healthcare institutions are securely transmitted to the cloud. There, diverse data undergoes harmonization, akin to orchestrating a symphony. Each institution's data contributes to a unified model, embodying collective knowledge and expertise. This collaborative effort showcases the power of shared wisdom from medical entities.

With aggregated training weights, our model makes accurate, nuanced predictions, informed by diverse medical data. This holistic approach integrates cloud computing, federated learning, and medical expertise, marking a new era in healthcare solutions. Facilitated by socket connections, it enables seamless communication, paving the way for data-driven, collective intelligence in healthcare decisions, enhancing patient care and outcomes (Fig. 5).

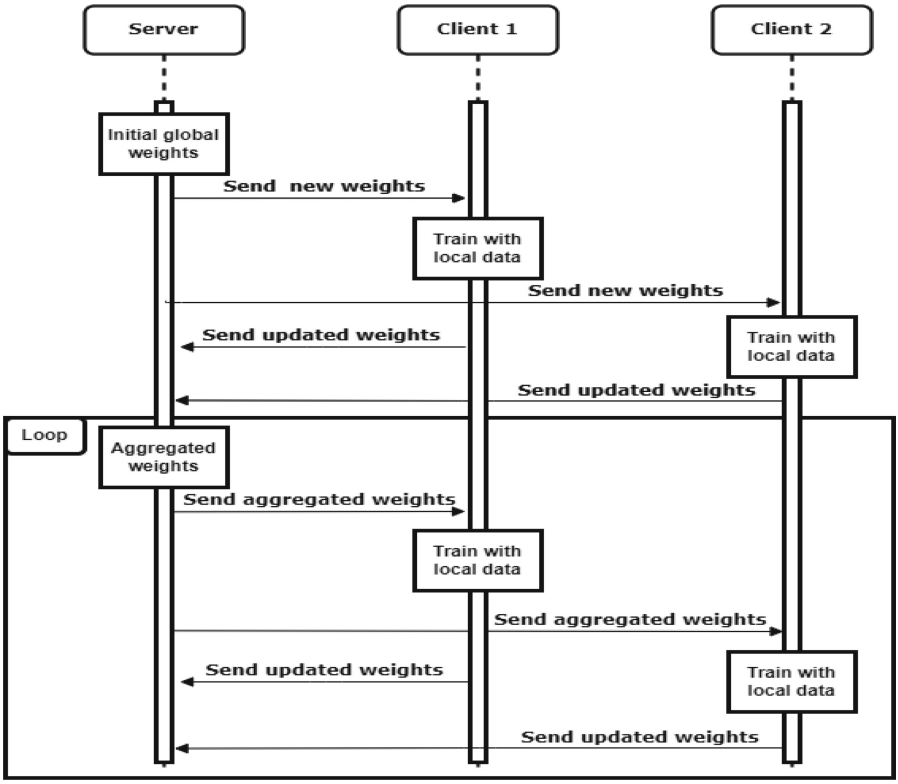

Fig. 5. Sequence diagram of Federated Learning.

3.4 Cluster K3S

Our healthcare solution utilizes K3S Clusters, powered by Raspberry Pi devices, to simulate hospital environments efficiently. Within these clusters, MySQL serves a dual role as a secure data repository and a client for our Federated Learning model. This integration enhances both data security and model capabilities, showcasing the sophistication of our healthcare solution (Fig. 6).

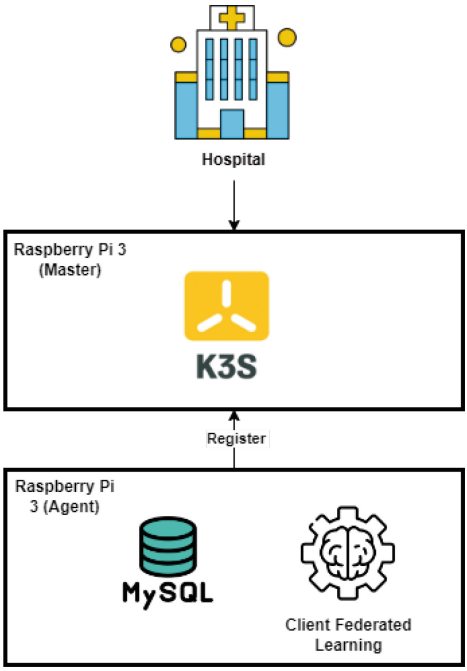

Fig. 6. Cluster K3S.

4 Performance Evaluation

4.1 Evaluation of Training Data

The dataset used for training in our study is the Pima Indian Diabetes (PIDD) dataset [11]. The dataset we are working with is fundamental to our diabetes prediction model, providing a rich source of information in CSV format. It consists of 768 rows and encompasses 9 essential columns, each capturing specific health parameters vital for our analysis.

Firstly, the 'Pregnancies' column documents the number of pregnancies each individual has undergone, offering valuable insights into their reproductive history. 'Glucose' stands for the fasting blood sugar levels, measured in milligrams per deciliter (mg/dL), providing critical information about glucose metabolism. The 'BloodPressure' column records diastolic blood pressure in millimeters of mercury (mm Hg), a key indicator of cardiovascular health.

Additionally, 'SkinThickness' measures the thickness of a skinfold at the triceps in millimeters (mm), offering data related to body composition. 'Insulin' signifies serum insulin levels, presented in micro-units per milliliter (μU/ml), providing insights into insulin production and regulation. 'BMI' or Body Mass Index, a derived numerical value, characterizes an individual's overall body composition, aiding in the assessment of weight-related health risks (Table 1).

The 'DiabetesPedigreeFunction' column quantifies the genetic influence of diabetes based on family history, shedding light on hereditary factors. 'Age' represents the individual's age in years, a critical factor in diabetes risk assessment (Table 2). Finally, the 'Outcome' column serves as a binary variable, taking the value '1' to indicate the presence of diabetes and '0' to signify its absence (Fig. 7).

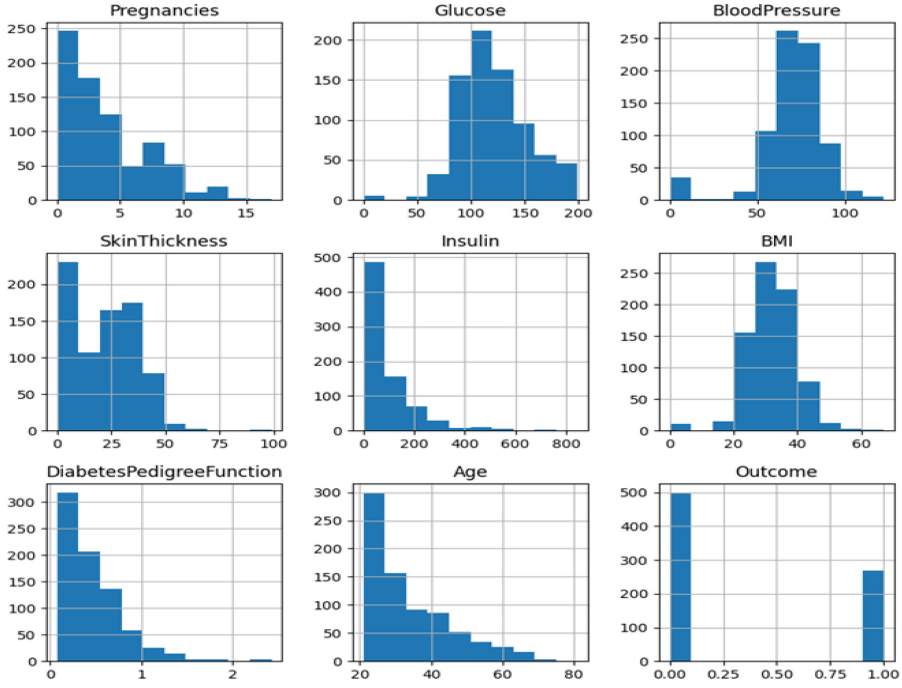

Fig. 7. Charts Representing Features.

Table 1. Description Distributed Data Statistics

Feature	Meaning
Count	the number of NoN-empty rows in a feature.
Mean	mean value of that feature.
Std	Standard Deviation Value of that feature.
Min	minimum value of that feature.
Max	maximum value of that feature.
25%, 50%, and 75%	are the percentile/quartile of each features

Table 2. Distributed Data Statistics

Count	768	768	768	768	768	768	768	768	768
Mean	3.84	120.89	69.1	20.53	79.79	31.09	0.47	33.24	0.34
Std	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76	0.47
Min	0	0	0	0	0	0	0.078	21	0
Max	17	199	122	99	846	67.1	2.42	81	1
25%	1	99	62	0	0	27.3	0.24	24	0
50%	3	117	72	23	30.5	32	0.37	29	0
75%	6	140.25	80	32	127.25	36.6	0.62	41	1

In our dataset analysis, a notable pattern emerged: as the number of pregnancies increased, insulin levels in women tended to decrease. This hints at a potential inverse correlation between pregnancies and insulin levels, especially in women. While promising, further detailed analysis and statistical confirmation are needed to validate this relationship's strength. If substantiated, this insight could significantly enhance our diabetes understanding, potentially leading to more accurate prediction models and focused prevention strategies. These preliminary findings inspire further research, driving us to explore deeper patterns and relationships, advancing our knowledge in diabetes prediction.

This finding is particularly crucial as it could have far-reaching implications for understanding the risk factors associated with diabetes. However, it's essential to note that further in-depth research and rigorous statistical analysis are necessary to confirm the strength and significance of this relationship. Despite the need for additional investigation, these initial observations serve as valuable and promising starting points for our ongoing exploration of patterns within the dataset.

Understanding such relationships can profoundly impact our ability to predict and prevent diabetes effectively. These preliminary insights provide a strong foundation for our future studies, guiding our focus as we delve deeper into the intricacies of the data, aiming to uncover more nuanced patterns and correlations for a comprehensive understanding of diabetes risk factors (Fig. 8).

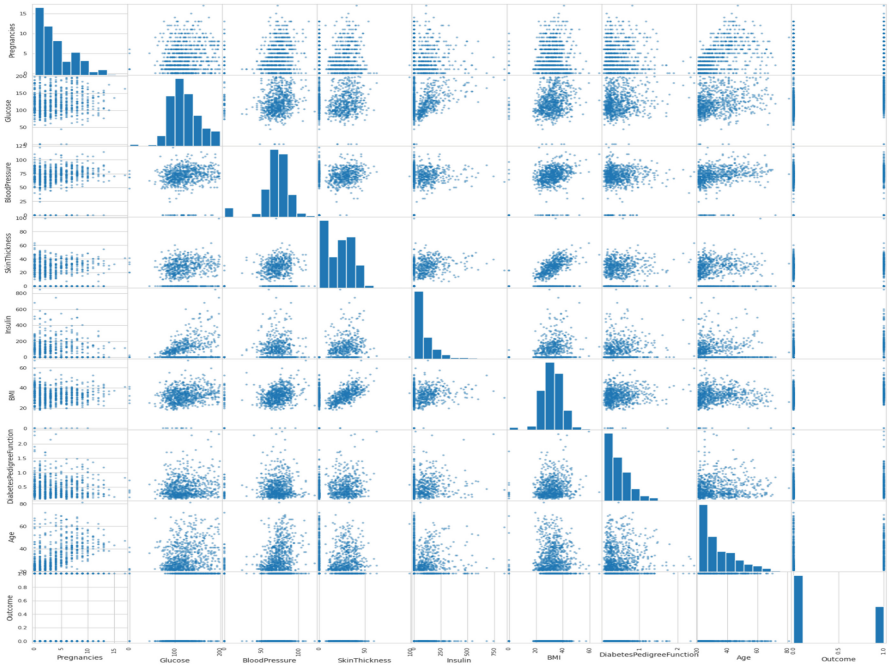

Fig. 8. Scatter Matrix

4.2 Evaluation of Predictive Models

In our environment, the server runs Ubuntu 22.04 on the OpenStack platform, while two Ubuntu 20.04 virtual machines on VMware 17 Pro serve as clients. The Pima Indian Diabetes (PIDD) dataset is divided into parts, with 60% residing on the server and 20% on each client. For aggregation, model weights are calculated as the average of weight files sent by each client to the server (Table 3).

Table 3. Performance Evaluation

Resources	Server	Client 1	Client 2
CPU (%)	57	31	30
RAM (MB)	1365	1090	1080

In terms of evaluation results, the original K-Nearest Neighbors (KNN) algorithm achieved an accuracy of 79%. However, after implementing Federated Learning (FL), the model's performance significantly improved to an accuracy of 87% (Table 4).

Table 4. Evaluation results of traditional KNN model and our proposal

Metric	KNN	KNN + FL	Description
Accuracy	0.79	0.87	Measures overall correctness
Precision	0.70	0.86	Gauges accurate positive predictions
Recall	0.63	0.81	Captures actual positive instances well
F1-Score	0.66	0.84	Balances precision and recall

These results suggest that FL has substantially enhanced the model's predictive capabilities in the described setting, making it a promising approach for improving diabetes prediction. Nonetheless, continuous refinement may be necessary to ensure that the model consistently achieves high accuracy while minimizing the risk of missing diabetes cases (Fig. 9).

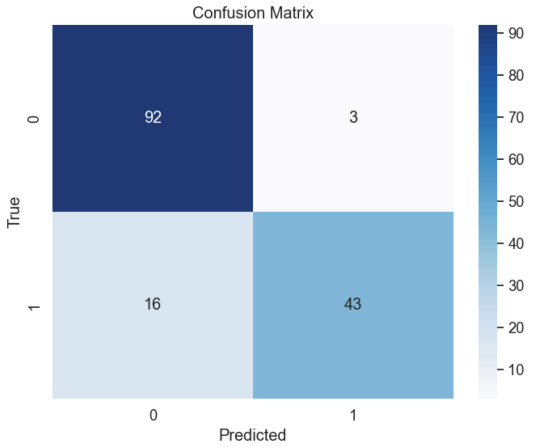

Fig. 9. Confusion Matrix.

5 Conclusion and Future Works

In this paper, we develop a solution for e-healthcare via a variety of technologies, such as Internet of Things, Federated learning, and Edge Computing. While IoTs and Edge Computing ensure the connectivity of users, devices, and services, Federated learning supports Deep Learning on learning users data without collecting all of them. The results show that our proposal outperform the centralized learning in deep learning evaluation metrics while keeping users data security and privacy.

The system offers several promising avenues for development. Firstly, it can expand its predictive capabilities beyond diabetes, encompassing other health-related conditions such as cardiovascular diseases and type 2 diabetes. This

expansion provides users with a more comprehensive view of their health. Secondly, ensuring the privacy and security of medical data is paramount, necessitating ongoing research into data protection measures, including data encryption and privacy management. Lastly, creating user-friendly mobile applications and interfaces is vital for enhancing user engagement and making it easy for individuals to track and manage their health information on-the-go.

Acknowledgment. This research is funded by University of Information Technology-Vietnam National University HoChiMinh City under grant number D1-2023–62

References

1. Alshamrani, M.: IoT and artificial intelligence implementations for remote healthcare monitoring systems: a survey. J. King Saud Univ. Comput. Inf. Sci. **34**(8), 4687–4701 (2022)
2. Smith, A.C., et al.: Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J. Telemedicine Telecare **26**(5), 309–313 (2020)
3. Li, Q., Liu, L., Zhou, Y.: A wearable remote monitoring system for chronic disease. IEEE Access **5**, 18309–18318 (2017)
4. Javaid, M., et al.: Significance of machine learning in healthcare: features, pillars and applications. Int. J. Intell. Netw. **3**, 58–73 (2022)
5. Khaleel, F.A., Al-Bakry, A.M.: Diagnosis of diabetes using machine learning algorithms. Mater. Today Proc. **80**, 3200–3203 (2023)
6. Al-Masni, M.A.N., Algani, N.: Prediction of diabetes mellitus using k-nearest neighbor algorithm based on feature extraction. IOP Conf. Ser. Mater. Sci. Eng. **459**(1), 012028 (2018)
7. Antunes, R.S., André da Costa, C., Küderle, A., Yari, I.A., Eskofier, B.: Federated learning for healthcare: systematic review and architecture proposal. ACM Trans. Intell. Syst. Technol. (TIST) **13**(4), 1–23 (2022)
8. Nguyen, D.C., et al.: Federated learning for smart healthcare: a survey. ACM Comput. Surv. (CSUR) **55**(3), 1–37 (2022)
9. Li, Y., Zhang, J., Liu, J., He, Y.: FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Trans. Industr. Inf. **16**(1), 188–197 (2020)
10. Avendano, J.P., et al.: Interfacing with the electronic health record (EHR): a comparative review of modes of documentation. Cureus **14**(6) (2022)
11. Yabo, M.M.I., Garko, A.B., Muslim, A.A., Suru, H.U.: A review of diabetes datasets. J. Comput. Sci. Appl. **10**(1), 6–15 (2022)
12. Dave, R., Seliya, N., Siddiqui, N.: The benefits of edge computing in healthcare, smart cities, and IoT (2021). *arXiv preprint*arXiv:2112.01250
13. Hartmann, M., Hashmi, U.S., Imran, A.: Edge computing in smart health care systems: review, challenges, and research directions. Trans. Emerg. Telecommun. Technol. **33**(3), e3710 (2022)
14. Saxena, R.: Role of K-nearest neighbour in detection of Diabetes Mellitus. Turk. J. Comput. Math. Educ. (TURCOMAT) **12**(10), 373–376 (2021)
15. Marfoq, O., Neglia, G., Vidal, R., Kameni, L.: Personalized federated learning through local memorization. In: International Conference on Machine Learning, pp. 15070-15092. PMLR (June 2022)
16. Liu, M., Ho, S., Wang, M., Gao, L., Jin, Y., Zhang, H.: Federated learning meets natural language processing: a survey (2021). *arXiv preprint arXiv*:2107.12603