Cloud Computing and Data Science
http://ojs.wiserpub.com/index.php/CCDS/

UNIVERSAL WISER PUBLISHER

Research Article

AI-Driven Federated and Transfer Learning Platform for Health Predictions

Venkatesh Upadrista[1*], **Alejandro Martinez Galindo**[2], **Murthy S**[3]

[1]Department of Computing, Glasgow Caledonian University, Glasgow G4 0BA, Scotland, United Kingdom
[2]Fortrea, Moore Drive, Durham, North Carolina 27709, United States of America
[3]Futurelight Technologies & Pfizer Joint Venture, United Kingdom
E-mail: vupadr200@caledonian.ac.uk

Received: 11 September 2024; **Revised:** 22 October 2024; **Accepted:** 5 November 2024

Abstract: Medical negligence, errors, and delayed diagnoses lead to preventable deaths and serious health issues. These problems are mainly caused by a shortage or lack of access to qualified healthcare professionals. In the U.S. alone, medical errors cause an estimated 44,000 to 98,000 deaths annually. The rise of artificial intelligence in healthcare has introduced new opportunities for more accurate diagnoses and predictions, with numerous artificial intelligence (AI)-driven tools now available for detecting diseases such as cancer, heart conditions, lung diseases, and liver diseases. While these tools have revolutionized diagnostics, they are primarily designed for healthcare professionals, limiting their accessibility to the general public. Providing patients with a reliable platform to help them accurately assess their health condition can be transformative, reducing medical errors, enhancing patient engagement, and improving overall care. This study introduces E-Doctor, a Generative AI-powered platform designed to provide reliable second opinions for medical diagnoses. Utilizing advanced AI technologies such as BioBERT for analyzing medical texts and federated learning for maintaining data privacy, the platform addresses the shortage of professional healthcare advice. E-Doctor platform achieved high accuracy rates: 92.4% for heart attack-related advice, 90.2% for full-body checkups, and 93.9% for Deep Vein Thrombosis prediction. By combining medical image analysis with AI-driven learning models, e-Doctor offers a robust platform for enhancing patient outcomes, while ensuring data privacy and security.

Keywords: predictive analytics, biomedical text analysis, patient data privacy, medical imaging AI, BioBERT, telemedicine system, AI healthcare platform, disease detection

1. Introduction

In many regions, the shortage of qualified medical professionals has resulted in increased medical negligence and delayed diagnoses. This leads to misdiagnoses, inadequate treatment, and a higher likelihood of medical errors, all negatively impacting patient outcomes.

In the U.S. alone, medical errors are a leading cause of death, accounting for an estimated 44,000 to 98,000 deaths annually [1]. In Germany, medical errors result in 25,000 deaths each year [2], and in Australia, they are responsible for 18,000 preventable deaths annually [2]. Preventable medical errors surpass deaths from major causes like accidents and cancer, emphasizing the importance of enhancing patient safety and reducing healthcare costs [1]. These alarming

DOI: https://doi.org/10.37256/ccds.6120255703

statistics highlight the critical need for safer healthcare systems and more accessible diagnostic tools to prevent adverse events and reduce the overwhelming burden on healthcare systems globally.

The rise of artificial intelligence (AI) in healthcare has created new opportunities for more accurate medical diagnoses and predictions [3]. Several innovative models have emerged, offering immense value to healthcare systems by improving disease detection and prediction. Numerous AI-driven tools are now available in the market, aiding in the diagnosis of various diseases such as cancer, heart conditions, lung diseases and liver diseases. For instance, Zebra Medical Vision [4] is an AI-powered medical imaging platform that assists radiologists in detecting diseases like cancer, cardiovascular conditions, and liver diseases using computed tomography (CT) scans, X-rays, and magnetic resonance imagings (MRIs). Another example is Tempus [5], an AI-driven platform that analyzes clinical and molecular data to help physicians identify genetic mutations linked to cancer and guide personalized treatment plans. Similarly, Arterys [6] is an AI tool designed for cardiovascular imaging, which helps in analyzing magnetic resonance imaging (MRI) and echocardiogram data to detect heart diseases, including heart failure and congenital heart conditions. In addition, various literature studies have developed applications for detecting different health conditions, such as brain stroke detection [7], COVID-19 infection detection [8] and post-stroke analysis [9, 10], which demonstrates the potential of AI in healthcare. While these AI tools and research have revolutionized diagnostics, they are primarily designed for physicians, limiting their accessibility and usability for the general public. Moreover, these tools typically focus on single pathologies, highlighting the need for a comprehensive AI-driven solution capable of detecting multiple conditions within a single platform.

The motivation for this study is rooted in addressing the significant gap between solutions available to healthcare professionals and the lack of tools empowering patients to understand their health conditions directly. Providing patients with a reliable tool to assess their health conditions and offer medical advice can be transformative, reducing medical errors and improving overall patient care. Harnessing the power of AI, such platforms can be developed to provide patients with accurate and timely second opinions, ensuring they receive the correct diagnosis and treatment options while enabling more effective engagement with their doctors. With a clearer understanding of their health issues, patients can ask informed questions and even validate their doctor’s diagnosis and treatment plan. This heightened awareness leads to better decision-making, increased trust in healthcare recommendations, and improved health outcomes.

This paper introduces e-Doctor, a platform that utilizes transfer learning and federated learning to analyze medical images and patient data, providing second opinions on various medical conditions. Generative AI (GenAI) technology was used to analyze medical images and patient information. Another key platform component is the BioBERT model, a version of Bidirectional Encoder Representations from Transformers (BERT) pre-trained on large biomedical datasets [11]. BioBERT excels in understanding medical literature and terminology, making it an ideal starting point for processing medical text [12, 13]. The model is further fine-tuned on disease-specific datasets, such as heart attack prediction, deep vein thrombosis (DVT) prediction, and general medical tests (full body check-ups). Fine-tuning BioBERT through transfer learning allows the model to adapt to specific tasks, enhancing its ability to make precise predictions based on the input medical text.

By integrating transfer learning, federated learning, and Gen AI technologies into a single platform, e-Doctor represents a major step forward in using AI for the well-being of individuals. The platform has demonstrated highly accurate medical advice across various conditions, achieving 92.4% accuracy for heart attack-related advice, 90.2% for full-body and urgent general checkups, and 93.9% for deep vein thrombosis (DVT) predictions. These results underscore its potential to improve patient outcomes while ensuring data privacy.

The remainder of the paper is structured as follows: Section 2 presents a literature review, while Section 3 outlines the proposed work. Section 4 covers the performance analysis. The experimental setup is described in Section 5, with datasets discussed in Section 6. Section 7 presents the results, and Sections 8 and 9 provide the discussion, conclusion, and directions for future research.

2. Literature review

AI in healthcare has experienced substantial growth in recent years, with applications ranging from diagnostic tools to personalized treatment plans. Machine learning models, especially transformer-based architectures like BERT, have shown exceptional performance in Natural Language Processing (NLP) tasks related to medical text analysis. Research also suggests that transfer learning, combined with fine-tuning on domain-specific datasets using NLP, significantly improves performance in specialized medical tasks such as cancer diagnosis, heart disease prediction, and DVT detection. Additionally, federated learning has emerged as a solution to the challenge of protecting data privacy in healthcare applications. This approach enables collaborative model training across institutions without the need to centralize sensitive patient data, thus addressing concerns about data security and privacy.

Our literature review focused on three key areas: BioBERT, transfer learning, and federated learning, each demonstrating significant advancements in healthcare applications.

Transfer learning has been applied across various medical contexts. DenseNet121, a pre-trained deep learning model, was fine-tuned to detect brain strokes achieving an accuracy of 97.45% [7]. Additionally, an IoT-based fuzzy logic framework predicted COVID-19 infection risks with the model reaching 80% accuracy [8]. Transfer learning was also used for hand gesture recognition in post-stroke patients, achieving 82.2% accuracy, significantly outperforming traditional classifiers like neural networks and light gradient-boosting machine (LGBM) [9]. Moreover, in another literature transfer learning was employed through session-specific strategies to enhance brain-computer interface (BCI) performance for stroke rehabilitation, with the instantaneous strategy reaching a classification accuracy of 76.4% [10]. BioBERT has shown exceptional results in biomedical tasks. For COVID-19-related named entity recognition (NER), BioBERT models achieved over 90% accuracy in identifying biomedical entities like diseases and genes [11-14]. In another example, BioBERT was fine-tuned to predict heart disease risk factors from electronic health records (EHR), achieving 93.99% accuracy [15]. Furthermore, BioBERT was utilized in medical imaging protocol classification, where it reached an F1 score of 0.92, improving the automation of neuroradiology tasks [16]. Federated learning has been widely applied in healthcare to address privacy concerns while enhancing prediction accuracy. This approach ensures patient data remains local, and only model updates are aggregated centrally. For instance, a federated learning framework for epileptic seizure detection allowed patients to train personalized models, achieving a sensitivity of 90.24% [17]. Another model inferred arterial blood pressure from wearable devices with a mean error of 2.95 mmHg [18]. Additionally, federated learning was applied to predict COVID-19 mortality using clinical data from multiple hospitals, outperforming local models [19]. In the detection of arrhythmia, federated learning minimized network traffic and energy consumption while maintaining high accuracy [20]. Finally, a federated learning driven IoMT framework for sleep monitoring achieved 74.19% accuracy, showing the potential of this technology in healthcare applications [21].

These research efforts collectively demonstrate the effectiveness of models like BioBERT, transfer learning, and federated learning in their respective applications. The review emphasizes that while BioBERT and transfer learning have shown substantial promise in analyzing medical texts and images, federated learning provides a powerful solution for improving model accuracy while preserving data privacy across institutions. However, none of these models are designed to detect multiple pathologies in one platform or offer patients suggestions regarding potential health risks. This creates a notable gap, especially as access to qualified medical professionals becomes more limited. A unified platform capable of addressing various medical conditions and offering reliable second opinions remains a critical unmet need, highlighting the potential for further innovation in this area that has led to the development of e-Doctor as part of this paper.

2.1 *Transfer learning*

Authors in [7], applied transfer learning using pre-trained deep learning models, specifically DenseNet121, ResNet50, and VGG16, to detect brain strokes from MRI images. These models are fine-tuned on brain MRI datasets to leverage their previously learned features for the specific task of stroke detection. The models achieved high accuracy, with DenseNet121 showing the best performance with an accuracy of 97.45%, followed by ResNet50 at 95.21%, and VGG16 at 91.12%. These results demonstrate the effectiveness of transfer learning in enhancing model performance for medical imaging tasks like brain stroke detection.

As part of this study, authors in [8] introduced an IoT and cloud-based architectural framework to predict the risk of COVID-19 infection using hesitant intuitionistic fuzzy sets (IFS) based on citizen self-assessments. The framework integrates self-assessment data and wearable IoT devices to predict infection risks and monitor patients remotely. The study focuses on fuzzy logic models and their applications in medical diagnosis. The model achieved an 80% accuracy for predicting infection based on real-time data.

Authors in [9] have applied transfer learning using prototypical networks for hand gesture recognition in post-stroke patients. The approach involves one-shot transfer learning, allowing models pre-trained on healthy individuals to adapt to new participants with limited training data. This method improves the accuracy of gesture classification from wearable sensors. The proposed model achieved an accuracy of 82.2%, outperforming other subject-independent classifiers such as neural networks (59.72%) and LGBM (65.09%). The results highlight the effectiveness of using few-shot transfer learning to address the variability in biological signals from stroke survivors.

In [10] transfer learning was employed through subject-specific session-to-session strategies to improve brain-computer interface (BCI) performance for stroke patients undergoing upper extremity rehabilitation. Three strategies were evaluated: the previous session, accumulative, and instantaneous strategies. The instantaneous strategy, which utilizes current and previous session data for training, showed the best performance, achieving a median classification accuracy of 76.4%. This strategy allows patients to adapt to variability across sessions, enhancing BCI control without reducing therapy time. The accumulative and previous session strategies were less effective but still contributed to performance improvements.

2.2 *BioBERT*

Authors in [11] discuss the application of the BioBERT model, fine-tuned on the COVID-19 Open Research Dataset (CORD-19), for named entity recognition (NER) tasks. The model helps in identifying and classifying biomedical entities related to COVID-19. Several pre-trained BioBERT models were used, and their performance was evaluated based on precision, recall, and F1-score. The BioBERT-Large v1.1 with PubMed 1 M showed superior results, with an F1-score of 0.93. The paper highlights BioBERT's capacity to enhance entity recognition in biomedical texts, significantly improving the extraction of meaningful information related to COVID-19.

Authors in [12] investigates the use of natural language processing (NLP) for automating abstract reviews in medical research, with a focus on the BioBERT model. BioBERT, a BERT-based model pre-trained on biomedical texts, was applied to classify 12,817 abstracts into three categories: text source, context of use, and primary research fields. The model achieved micro F1-scores of 77.35%, 76.24%, and 85.64%, respectively, demonstrating robust classification performance.

In [13], the BERT model was used to correct substitution errors in Mandarin Chinese speech recognition. The authors propose EC-BERT, a BERT-based error correction model that directly learns from context to identify and correct errors without relying on mask mechanisms or error detection networks. EC-BERT is pre-trained with pseudo-paired data and fine-tuned with real data, improving ASR performance. Experimental results show that EC-BERT reduces character error rates by 19.2% compared to CTC Greedy Search and 12.8% compared to CTC-WFST results, while also demonstrating faster processing than similar models.

In [14] BioBERT was used to extract clinical and demographic information from COVID-19 case reports. The fine-tuned BioBERT model achieves a named entity recognition accuracy improvement of 1-3% over benchmark methods and an F1 score of 90.58% for clinical tasks. This work demonstrates how transfer learning enhances information extraction and relation prediction without extensive labeled data, aiding public health surveillance.

Authors in [15] explored adapting transformer-based models, including BioBERT, for heart disease detection and extraction of risk factors from Electronic Health Records (EHRs). They used transfer learning techniques to fine-tune pre-trained models such as BERT, BioClinicalBERT, XLNet and BioBERT on heart disease-specific datasets. The models demonstrated superior performance in identifying risk factors like diabetes, hypertension, and smoking. The model achieved the highest performance, with a micro F1-score of 94.27%, while other models like BioBERT achieved 93.99%.

Authors [16] fine-tuned various BERT-based models, including BioBERT, ClinicalBERT, and RoBERTa, for medical imaging protocol classification. They focused on improving the performance and explainability of these models by assessing gradient-based methods. The study aimed to automate neuroradiology protocol assignment, reducing the workload on radiologists. Transfer learning played a key role in enhancing model accuracy. The experimental results showed that BioBERT achieved an F1 score of 0.92, outperforming other models.

2.3 *Federated learning*

According to authors in [17] sending raw patient data to a centralized cloud server not only puts patient privacy at risk but it consumes a lot of energy. To overcome this challenge authors designed and evaluated a standard Federated learning framework for epileptic seizure detection using a deep learning-based local battery-powered mobile platform and the server. The model achieved a sensitivity of 81.25%, a specificity of 82.00%, and a geometric mean of 81.62%. A customized form of Federated learning was also examined by the authors which provided an improvement to the selected performance metrics, with a sensitivity of 90.24%, a specificity of 91.58%, and a geometric mean of 90.90%.

Authors in [18] developed a machine learning framework that can estimate arterial blood pressure (ABP) using just a single optical photoplethysmogram (PPG) sensor. They trained the model using multiple distributed local models and data sources, simulating a large-scale collaborative learning environment suitable for low-cost wearable devices. The results showed an average error of 2.95 mmHg with a standard deviation of 19.33 mmHg. This framework lays the groundwork for continuous large-scale ABP monitoring by providing a sustainable, privacy-preserving real-world solution. It also enables more power-efficient learning, allowing the development of smaller, more affordable wearable devices without compromising the accuracy of ABP measurements or patient privacy.

In [19], the authors assessed the effectiveness of using Federated Learning to predict mortality in COVID-19 patients by analyzing clinical data from Electronic Health Records (EHR). Data from five hospitals, each with different patient demographics, outcomes, sample sizes, and lab values, were used to simulate a real-world scenario with diverse patient populations. The study demonstrated that Federated Learning applied to both Multilayer Perceptron (MLP) and Logistic Regression with L1 regularization (LASSO) models, outperformed the individual local models at each hospital. These results suggest that Federated Learning can help overcome the limitations of isolated local models by combining insights across different hospitals, leading to better predictive accuracy.

Authors in [20] proposed an advanced federated learning framework to train deep neural networks for detecting arrhythmia by monitoring single-lead electrocardiograms. The network was partitioned into segments and allocated across multiple IoMT devices for federated learning, with a portion assigned to a centralized server. This model ensured that most of the computational load was managed by a powerful centralized server, rather than relying on the constrained edge devices. Experiments demonstrated minimal accuracy loss, with only 0.2% of the synchronization traffic required by the vanilla federated learning approach.

Authors in [21] propose a Federated Learning Driven Internet of Medical Things (FLDIoMT) framework to enhance privacy and security in IoMT applications. The framework enables flexible deployment of medical services while safeguarding sensitive patient data. They introduce a novel sleep monitoring system, iSmile, as a practical application of the framework, demonstrating its effectiveness in real-world use. The results from their simulated experiment show that the federated model, trained across multiple hospitals, achieved a model accuracy of 74.19%, which closely matches the 77.42% accuracy of a centralized model, validating the feasibility of their approach.

Authors in [22] proposed a novel approach combining federated learning and Generative Adversarial Networks (GANs) to enhance healthcare systems through privacy-preserving and secure model training. The technique, called Customized Inequality-Aware Federated Learning (CusIAFL), improves communication security and model performance in wireless networks. The model achieved 89-93% accuracy on various medical imaging datasets, demonstrating its potential to create robust and resilient healthcare systems while safeguarding patient data privacy.

Authors in [23] focused on classifying URLs to detect malicious content using machine learning and deep learning techniques. They employed models like Random Forest, Decision Trees, and CatBoost for classical machine learning models, while also exploring several deep neural networks. The best-performing deep learning model achieved an accuracy of 95.61%. This study demonstrates the effectiveness of both classical and deep learning models in identifying malicious URLs, with feature extraction and engineering playing a significant role in improving model performance.

Authors in [24] explored the intersection of AI and Human-Computer Interaction (HCI) across various applications. They emphasized how AI is transforming fields such as education, healthcare, and business by enhancing user experiences and optimizing systems through AI-powered interfaces. Key areas include augmented reality, AI explainability, and human-centered AI. Various case studies showcase the implementation of intelligent user interfaces and the use of AI for predicting employability and handling tasks in healthcare, public security, and smart systems.

3. Proposed work

This study introduces e-Doctor, an AI-powered platform designed to offer patients reliable second opinions based on their health conditions and medical reports. It addresses the critical shortage of qualified healthcare professionals and the growing concerns regarding medical errors. Unlike existing AI tools, which are primarily geared toward healthcare professionals, e-Doctor empowers patients to interact directly with their health data. By utilizing advanced technologies such as transfer learning, federated learning, and models like BioBERT, the platform delivers accurate and secure medical advice across various health conditions. The platform is user-friendly; patients simply enter their information, describe their health concerns, and upload medical records. Based on this data, e-Doctor provides personalized recommendations. The proposed platform ensures accurate diagnostic predictions while safeguarding patient privacy.

3.1 *Key technologies used*

Transfer Learning for Medical Predictions: The platform utilizes transfer learning to enhance its prediction capabilities for various medical conditions. A pre-trained model is further fine-tuned on disease-specific datasets to provide precise predictions for conditions like heart attacks, deep vein thrombosis (DVT), and overall health evaluations. This approach allows the platform to achieve high diagnostic accuracy with minimal training data, reducing the complexity and cost of creating disease-specific models from scratch.

Federated Learning for Data Privacy: To ensure patient data privacy and meet regulatory standards like the Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR), the platform employs federated learning. This decentralized model allows patient data to remain local within healthcare institutions while training a global model. Instead of sharing raw patient data, only the model's updated parameters are communicated and aggregated to improve prediction capabilities. This ensures that the platform provides accurate medical insights without compromising data privacy.

NLP-Powered Diagnostics Using BioBERT: At the heart of the platform's diagnostic capabilities is the BioBERT model, specifically fine-tuned for analyzing medical records and health data. BioBERT's deep learning-based architecture, trained on vast biomedical texts, allows it to process complex medical terminology and provide insights based on unstructured text, including electronic health records (EHRs), patient symptoms, and diagnostic reports. This makes it well-suited for predictive tasks, such as early detection of diseases and interpreting medical images.

3.2 *Expected outcomes*

The e-Doctor platform is expected to improve patient engagement in managing their health by providing actionable insights and second opinions. The key outcomes of the proposed system include:

- Accurate and Reliable Medical Predictions: The platform emphasizes high accuracy and can provide health recommendations and advice for heart conditions, DVT, and general health assessments.
- Scalable and Secure AI Platform: By using federated learning, the system ensures data privacy and security while maintaining scalability for future healthcare applications.
- Improved Patient Empowerment: Empowering patients to have access to reliable second opinions on their health status can reduce medical errors, facilitate better communication with healthcare providers, and promote informed decision-making.

In summary, this proposed work aims to create an innovative AI platform that bridges the gap between advanced medical diagnostics and patient access, ensuring secure, scalable, and high-accuracy medical predictions while safeguarding patient privacy.

4. Performance analysis

The performance analysis of the e-Doctor platform focuses on evaluating its predictive accuracy across various medical conditions. The platform was tested with datasets for heart attack prediction, full-body checkups, and deep vein thrombosis, showing high accuracy in all tasks.

• Heart Attack Prediction: The model achieved a 92.4% accuracy, with a precision of 91.1% and recall of 89.7%, showcasing its effectiveness in predicting cardiovascular conditions.

• Full-Body Checkups and Urgent General Checkups: The platform demonstrated 90.2% accuracy, underscoring its capability to evaluate general health and provide valuable insights for comprehensive medical assessments.

• DVT Prediction: For DVT, the platform exhibited an accuracy of 93.9%, reflecting its strength in diagnosing this condition.

These high accuracy levels are attributed to the use of BioBERT, which was specifically fine-tuned on biomedical texts and condition-specific datasets. This customization enhanced the model's ability to understand complex medical language, ensuring precise and reliable predictions.

5. Experimental set-up

Our experiment aimed to evaluate the performance of the e-Doctor platform across multiple medical prediction tasks using various techniques. The setup involved:

• Integration of Medical AI Tools: We successfully integrated proprietary medical AI tools to analyze medical images such as ECGs, MRI scans, and X-rays. Amazon Textract was employed to automatically extract text and data from scanned documents. Additionally, we utilized Zebra Medical Vision, which offers AI-powered medical imaging analysis for X-rays, CT scans, and MRI scans. Zebra Medical's AI algorithms are capable of detecting conditions like fractures, lung abnormalities, and cardiovascular issues. The system automatically generates reports based on these analyses, and the API is invoked using AWS services via Lambda functions.

• BioBERT Model: We utilized the BioBERT model, pre-trained on large biomedical corpora such as PubMed and PMC. BioBERT is specifically designed to understand medical language and terminology. We further fine-tuned this model on specialized, disease-specific datasets to enhance its predictive capabilities for conditions such as heart attacks, cancer, and DVT.

• Transfer Learning: Using transfer learning, we fine-tuned the BioBERT model with data specific to particular diseases. By transferring knowledge from the pre-trained model to these specialized datasets, we improved the model's performance in diagnosing and predicting disease outcomes, ensuring that it adapts to specific medical conditions like heart attacks and DVT.

• Federated Learning: To ensure the privacy and security of sensitive patient data, we implemented federated learning. Instead of transferring patient data to a central server, the model was trained locally at each healthcare institution using their own data. Only encrypted model updates, such as weight adjustments, are shared with a central server, which aggregates these updates to improve the global model. Most importantly, no raw data is exchanged, ensuring compliance with privacy regulations like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). By keeping patient data localized and encrypting communications, federated learning greatly reduces the risk of data breaches and unauthorized access, offering strong security while benefiting from collective insights across distributed datasets. A detailed discussion of the federated learning architecture can be found in Appendix 1.

6. Datasets used

To train and evaluate the performance of our models, we utilized the following datasets:

• Heart Attack Prediction Dataset [25]: This is a public dataset consisting of 5,000 patient records with annotated symptoms and diagnoses of heart attacks. Refer to Appendix 2 for more details on the datasets.

• Cancer Prediction Dataset [26]: A dataset comprising 7,500 records covering various types of cancer, including symptoms and outcomes. Refer to Appendix 3 for more details on the datasets.

• DVT Dataset [27]: This dataset, containing 4,000 records, was used for predicting and diagnosing DVT. Refer to Appendix 4 for more details on the datasets.

• Fully body checkup and urgent general checkup [28]: This dataset consists of 6,500 patient records that cover comprehensive medical evaluations, including routine full-body checkups and urgent general checkups for immediate medical concerns. Refer to Appendix 5 for more details on the datasets.

6.1 *Test data*

To simulate patient data for machine learning model training and testing, three files were used: “Patient_EHR.csv”, “Patient_SuppData.csv”, and “Patient_RealData.csv”. Each file contains different aspects of fictional patient data to simulate real-world medical scenarios.

Patient Electronic Health Record (EHR): This file contains long-term patient health data, including information about heart disease, cancer, diabetes, and stroke history. The machine learning algorithm uses Age, Sex, Past History of Heart Disease, and Past History of Stroke to predict different health conditions, while the other data fields assist in further diagnoses. Table 1 below shows the structure and example content of the Patient Electronic Health Record (EHR) file.

Table 1. Patient EHR file

Patient ID	Patient name	Age	Sex	Heart disease	Cancer disease	Diabetic	Stroke
1	Patient 1	32	Male	Yes	Yes	Yes	Yes
2	Patient 2	25	Male	No	No	No	No
3	Patient 3	70	Male	No	No	No	No
4	Patient 4	66	Female	Yes	Yes	Yes	Yes
5	Patient 5	89	Female	No	No	Yes	Yes
6	Patient 6	22	Female	No	No	No	No
7	Patient 7	54	Male	No	No	No	No

Disclaimer: All data presented in this table and document is entirely fictional and has been created solely for illustrative purposes. It is not derived from real patient records and adheres to data protection regulations, including GDPR and HIPAA. No personal or sensitive information related to actual individuals has been used or disclosed in this material

Table 2. Patient supplementary data file

Patient ID	Patient name	Work type	Residence type	Smoking status	Pollution levels	Married
1	Patient 1	Govt. Job	Rural	Yes	High	Yes
2	Patient 2	Software	Rural	No	Low	No
3	Patient 3	Retired	Rural	No	Normal	Yes
4	Patient 4	Retired	Rural	No	High	Yes
5	Patient 5	Retired	Rural	Yes	High	Yes
6	Patient 6	Software	Rural	No	Normal	Yes
7	Patient 7	Software	Rural	No	Normal	Yes

Disclaimer: All data presented in this table and document is entirely masked and has been created solely for illustrative purposes. It is not derived from real patient records and adheres to data protection regulations, including GDPR and HIPAA. No personal or sensitive information related to actual individuals has been used or disclosed in this material

Patient Supplementary Data: This file includes additional information about patient's lifestyle, residence, and environmental conditions. These fields are used by the machine learning algorithm to improve prediction accuracy for different health conditions. Data fields such as Work Type, Residence Type, Smoking Status, and Pollution Levels provide further context for assessments. Table 2 illustrates the structure and example content of the Patient Supplementary Data file.

Patient Transactional Data: This file contains real-time, transactional data that includes medical parameters such as Body Mass Index (BMI), Resting Average Blood Pressure, Average Cholesterol Levels, and Average Glucose Levels. Table 3 presents the structure and example content of the Patient Transactional Data file.

Table 3. Patient transactional data file

Date	Time	Patient ID	BMI	Resting average BP	Average cholesterol	Avg fasting blood	Max heart rate	Exercise-induced	Hypertension	Avg glucose level	Resting ECG	Old peak	ST slope	Chest pain type
16/12/21	15:20	1	146	150	130	100	155	No	120	150	Normal	120	Flat	TA
16/12/21	15:20	2	145	145	156	165	154	No	121	156	Normal	123	Flat	TA
15/12/21	16:12	3	150	150	160	170	161	No	130	180	Normal	150	Flat	TA
18/01/21	17:20	3	155	155	165	175	166	No	135	185	Normal	155	Flat	TA
18/03/21	19:20	4	147	147	157	167	158	Yes	127	177	High	147	Up	ATA

Disclaimer: All data presented in this table and document is entirely masked and has been created solely for illustrative purposes. It is not derived from real patient records and adheres to data protection regulations, including GDPR and HIPAA. No personal or sensitive information related to actual individuals has been used or disclosed in this material

These three data files are stored in a PostgreSQL database to enable efficient querying and retrieval. The metadata is securely stored using a Ganache private blockchain to ensure the integrity and confidentiality of the data.

7. Results

The experimental results demonstrated the platform's ability to achieve high accuracy across multiple medical prediction tasks. Table 4 provides a summary of the performance metrics for each task.

Table 4. e-Doctor results

Task	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Heart attack prediction	92.4	91.1	89.7	90.4	0.93
Full body checkup/Urgent general checkup	90.2	88.5	87.3	87.9	0.91
Deep vein thrombosis prediction	93.9	86.7	85.4	86.0	0.90

The results from Table 4 indicate that the e-Doctor platform achieved high predictive accuracy across all tasks. Specifically, for heart attack prediction, the model attained an accuracy of 92.4%, with a precision of 91.1% and recall of 89.7%. The model's Area Under the Curve (AUC) of 0.93 further emphasizes its high sensitivity and specificity in predicting cardiovascular conditions.

For Full-Body Checkups and Urgent General Checkups, the platform achieved a 90.2% accuracy, which demonstrates its ability to provide accurate and actionable insights in general health evaluations. The F1-score of 87.9% shows a balanced performance across precision and recall, making it effective in delivering accurate diagnoses even in routine medical checkups.

The platform performed exceptionally well in Deep Vein Thrombosis (DVT) prediction, achieving an accuracy of 93.9%. The high accuracy and F1-score of 86.0% reflect its robustness in handling conditions with lower prevalence but higher risks.

These high accuracy rates were achieved due to the combination of advanced AI models, including BioBERT, which was fine-tuned using disease-specific datasets. This fine-tuning enhanced the model's understanding of complex medical terminology, contributing to reliable and accurate predictions across diverse medical conditions. Moreover, the application of transfer learning improved model adaptability, while federated learning ensured that privacy standards were maintained without compromising prediction accuracy.

The integration of federated learning led to only a minimal 1-2% performance drop compared to centralized models, as seen in these tasks. This performance degradation is an acceptable trade-off for ensuring data privacy, especially in healthcare scenarios where data sensitivity is critical.

Table 5 below presents a comparison of the e-Doctor results with the various models discussed in the literature.

Table 5. e-Doctor results comparison with literature reviews

Task	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Reference
Heart attack prediction	92.4	91.1	89.7	90.4	0.93	e-Doctor platform
Full body/Urgent general checkup	90.2	88.5	87.3	87.9	0.91	e-Doctor platform
Deep vein thrombosis (DVT)	93.9	86.7	85.4	86.0	0.90	e-Doctor platform
COVID-19 infection prediction	80	-	-	-	-	Transfer learning and IoT-based fuzzy logic [8]
Post-stroke gesture recognition	82.2	-	-	-	-	Transfer learning (Prototypical networks) [9]
COVID detection (NER)	93	-	-	-	-	BioBERT [11, 12, 14]
Medical imaging protocol classification	92 (F1 score)	-	-	92 (F1 score)	-	BioBERT [16]
Epileptic seizure detection	90.24	90.24 (Sensitivity)	-	-	-	Federated learning [17]

Contrasting e-Doctor's performance with the literature results as demonstrated in Table 5, it is evident that the platform has demonstrated increased accuracy as compared to other existing models in the literature while ensuring data privacy. The results from the e-Doctor platform underscore the system's robustness in medical advice tasks achieving high accuracy across all medical conditions.

8. Discussion

The e-Doctor platform has demonstrated its effectiveness in delivering accurate medical predictions while preserving patient privacy through federated learning. By integrating transfer learning with the BioBERT model, the platform exhibits flexibility and adaptability across a variety of medical conditions. The results show strong performance across multiple tasks. For heart attack prediction, the platform achieved an accuracy of 92.4%, with a precision of 91.1%, recall of 89.7%, F1-score of 90.4%, and an AUC of 0.93, indicating high predictive capability. Similarly, for full body checkups and urgent general checkups, the platform achieved an accuracy of 90.2%, precision of 88.5%, recall of 87.3%, F1-score of 87.9%, and an AUC of 0.91, demonstrating robust general health assessment abilities. The DVT prediction task performed well with an accuracy of 93.9%, precision of 86.7%, recall of 85.4%, F1-score of 86.0%, and an AUC of 0.90.

While these results indicate high predictive accuracy, there is room for improvement, particularly in fine-tuning the model for more complex medical conditions and expanding the scope by incorporating additional medical datasets. The use of federated learning for decentralized model training has proven to be a viable solution, enabling collaboration between institutions without compromising patient privacy. Notably, the use of federated learning did not significantly hinder model performance, making it an ideal approach for medical applications where data security and privacy. The platform's ability to balance high performance with data privacy highlights its potential as a scalable platform for enhancing medical diagnostics.

9. Conclusion and future work

The e-Doctor platform has demonstrated significant potential in providing accurate and reliable medical predictions by integrating advanced AI technologies such as transfer learning, federated learning, and the BioBERT model. The experimental results demonstrate that the platform achieved high accuracy rates across various medical tasks, including 92.4% for heart-related advice, 90.2% for full-body checkups, and 93.9% for DVT-related advice, all while ensuring patient data privacy through federated learning. The model's ability to offer robust performance across multiple conditions, combined with its privacy-preserving capabilities, highlights its strength in addressing both prediction accuracy and security in healthcare. The benefits of e-Doctor extend beyond accuracy. By providing reliable second opinions, the platform empowers patients with actionable insights, allowing them to approach healthcare professionals with a clearer understanding of their conditions. This fosters better patient-doctor communication, promotes trust in healthcare recommendations, and improves overall patient outcomes. The implication of this study suggests that platforms like e-Doctor can play a crucial role in modern healthcare, particularly in regions with limited access to qualified medical professionals. As the model continues to evolve and expand its capabilities, it holds the potential to revolutionize the delivery of healthcare, making high-quality, AI-driven medical diagnostics more accessible and secure.

As part of our future work, we aim to extend the system's capabilities to detect additional conditions, including lung diseases, brain tumors, kidney failure, multiple sclerosis, and Parkinson's disease. Expanding to conditions like lung diseases and multiple sclerosis presents challenges due to the complexity and variability of medical data, such as imaging quality and diverse data sources. To address these challenges, we will use transfer learning to leverage existing knowledge and apply multi-modal learning to integrate various data types, including imaging, clinical records, and genetics. Advanced architectures like CNNs for imaging and RNNs for time-series data will be employed to improve adaptability.

Authors' contributions

All authors read and approved the final manuscript. The corresponding author was responsible for the study's conception and design. Venkatesh Upadrista authored the first draft of the manuscript. Several prominent medical professionals, including the chairman of Futurelight Technologies and the Chief Information Officer of Fortrea, reviewed the manuscript (as listed in the authorship). All authors have read and approved the final version of the manuscript.

Funding

Futurelight Technologies has providing the hosting environments, and no other funding was received as part of this

project.

Data availability

The datasets generated during and/or analyzed during the current study are confidential and not disclosed as part of this paper.

Conflict of interests

The authors have no relevant financial or non-financial interests to disclose.

References

[1] Kohn LT, Corrigan JM, Donaldson MS. Errors in health care: a leading cause of death and injury. *To Err is Human: Building a Safer Health System*. US: National Academies Press; 2000.

[2] Sheikhtaheri A, Sadeqi-Jabali M, Hashemi-Dehaghi Z. Physicians' perspectives on causes of health care errors and preventive strategies: a study in a developing country. *Iranian Journal of Public Health*. 2018; 47(5): 720.

[3] Sakib S, Fouda MM, Fadlullah ZM, Abualsaud K, Yaacoub E, Guizani M. Asynchronous federated learning-based ECG analysis for arrhythmia detection. In: *2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*. Athens, Greece: IEEE; 2021. p.277-282.

[4] Weber M, Bauerschmidt L. *Equity Research on Zebra Technologies Corp-Accelerated Digitization Along Supply Chains Promises Further Growth*. 2021. Available from: https://www.zebra.com/ [Accessed 6 November 2024].

[5] Tempus. *AI-Enabled Precision Medicine*. 2020. Available from: https://www.tempus.com/ [Accessed 6 November 2024].

[6] Arterys. 2024. Available from: https://www.arterys.com/ [Accessed 6 November 2024].

[7] Polamuri SR. Stroke detection in the brain using MRI and deep learning models. *Multimedia Tools and Applications*. 2024; 1-18.

[8] Tyagi NK, Tyagi K. IoT and cloud-based COVID-19 risk of infection prediction using hesitant intuitionistic fuzzy set. *Soft Computing*. 2024; 28(5): 3743-3755.

[9] Sarwat H, Alkhashab A, Song X, Jiang S, Jia J, Shull PB. Post-stroke hand gesture recognition via one-shot transfer learning using prototypical networks. *Journal of NeuroEngineering and Rehabilitation*. 2024; 21(1): 100.

[10] Carino-Escobar RI, Franceschi-Jimenez LA, Carrillo-Mora P, Cantillo-Negrete J. Subject-specific session-to-session transfer learning strategies for increasing brain-computer interface performance during upper extremity neurorehabilitation in stroke. *Journal of Medical and Biological Engineering*. 2024; 44(4): 596-606.

[11] Soni G, Verma S, Sharan A, Ahmad O. BioBERT-based model for COVID-related named entity recognition. In: *International Conference on Advances in IoT and Security with AI*. Singapore: Springer; 2023. p.333-346.

[12] Masoumi S, Amirkhani H, Sadeghian N, Shahraz S. Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. *Systematic Reviews*. 2024; 13(1): 107.

[13] Xiao S, Hao R, Cheng G, Xu X, Li T. EC-BERT: A BERT language model with error correction for mandarin Chinese speech recognition. *Journal of Shanghai Jiaotong University (Science)*. 2024; 1-7.

[14] Raza S, Schwartz B. Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach. *BMC Medical Informatics and Decision Making*. 2023; 23(1): 20.

[15] Houssein EH, Mohamed RE, Hu G, Ali AA. Adapting transformer-based language models for heart disease detection and risk factors extraction. *Journal of Big Data*. 2024; 11(1): 47.

[16] Talebi S, Tong E, Li A, Yamin G, Zaharchuk G, Mofrad MRK. Exploring the performance and explainability of fine-tuned BERT models for neuroradiology protocol assignment. *BMC Medical Informatics and Decision Making*. 2024; 24(1): 40.

[17] Baghersalimi S, Teijeiro T, Atienza D, Aminifar A. Personalized real-time federated learning for epileptic seizure detection. *IEEE Journal of Biomedical and Health Informatics*. 2021; 26(2): 898-909.

[18] Brophy E, De Vos M, Boylan G, Ward TE. Estimation of continuous blood pressure from ppg via a federated learning approach. *Sensors*. 2021; 21(18): 6311.

[19] Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR Medical Informatics*. 2021; 9(1): e24207.

[20] Yuan B, Ge S, Xing W. A federated learning framework for healthcare iot devices. *arXiv:2005.05083*. 2020. Available from: https://doi.org/10.48550/arXiv.2005.05083.

[21] Fan J, Wang X, Guo Y, Hu X, Hu B. Federated learning driven secure internet of medical things. *IEEE Wireless Communications*. 2022; 29(2): 68-75.

[22] Murmu A, Kumar P, Moparthi NR, Namasudra S, Lorenz P. Reliable federated learning with GAN model for robust and resilient future healthcare system. *IEEE Transactions on Network and Service Management*. 2024; 21(5): 5335-5346.

[23] Rayyan A, Aburas MG, Al-Mousa A. Uniform resource locator classification using classical machine learning & deep learning techniques. *Cloud Computing and Data Science*. 2023; 4(1): 17-30.

[24] Khan SB, Chandna S, Xhafa F, Namasudra S, Mashat A. *Innovations in Artificial Intelligence and Human-Computer Interaction in the Digital Era*. Elsevier; 2023.

[25] Rahman R. *Heart Attack Analysis & Prediction Dataset*. Kaggle. 2020. Available from: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset [Accessed 15 September 2024].

[26] Suwal MS. *Breast Cancer Prediction Dataset*. Kaggle. 2018. Available from: https://www.kaggle.com/datasets/merishnasuwal/breast-cancer-prediction-dataset [Accessed 12 October 2024].

[27] Biobank UK. *Non-Cancer Illness Code Self-Reported: Deep Venous Thrombosis (dvt)*. Medical Research Council Integrative Epidemiology Unit (MRC IEU). 2017. Available from: https://gwas.mrcieu.ac.uk/datasets/ukb-a-65/ [Accessed 12 October 2024].

[28] Santiago J. *CDC's Complete Dataset of 2018 in the 500 Cities Project, Now PLACES*. Kaggle. 2021. Available from: https://www.kaggle.com/datasets/jennifersantiago/500-cities-local-data-for-better-health-2018 [Accessed 12 October 2024].

Appendix 1: Federated learning architecture

Federated learning ensures that sensitive patient data remains localized, never leaving the device or institution where it is collected. Instead of transferring raw data, only the model updates (such as weight adjustments) are shared. This enables the development of powerful machine learning models without compromising privacy. Additionally, the architecture ensures that models trained across distributed datasets are tamper-resistant, preventing the tracing of data back to individual patients. FL also aggregates knowledge from multiple decentralized datasets, allowing for a more accurate and comprehensive model while adhering to privacy regulations like HIPAA and GDPR.

A scalable Federated learning architecture comprises of three layers as shown in Figure A1. The foundation of federated learning is quite simple. Every client that has data with them, including smartphones, hospitals, patients, and medical facilities, trains their unique models using their own data. They then send the model (not the data) to a centralized server, which combines them and distributes the updated combined model to each client for additional updating cycles.

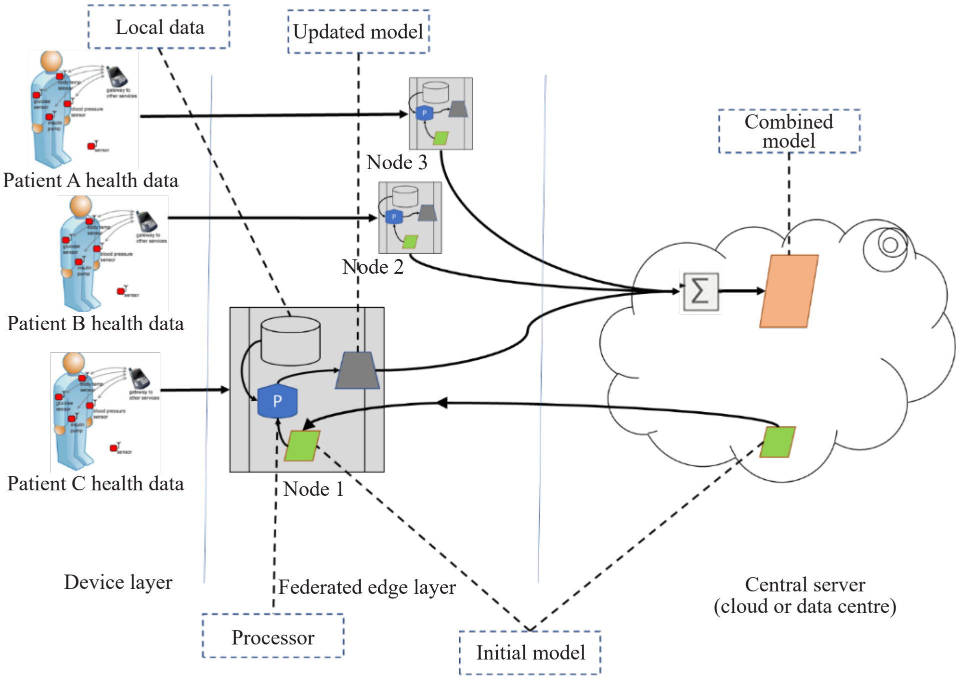

Figure A1. Federated learning data flow

Device layer

Connected healthcare devices provide data for the Internet of Medical Things. The physical parameters from the real world, such as patients health data, can be acquired using different sensors. The healthcare devices are composed of hardware and software, with some modern devices with on-board operating system, processing, storage, and connectivity. Some examples of Internet of Medical Things (IoMT) devices are Blood Pressure devices, Heartrate monitoring devices, telemetry device, and activity trackers. Other devices, such as insulin pumps and infusion are most prominently used in remote wellness and chronic disease monitoring. Many devices have the capability to store only limited data and perform data processing, however, due to the large data size generated by IoMT devices, data processing at device layer is almost impossible for healthcare use cases.

Federated edge layer

The processing of data collected by the IoMT devices takes place at this layer. The local model can improve in performance by learning from multiple datasets, and at the same time ensuring that data of one entity is not shared with others, thereby preserving privacy.

As the first step a base machine learning model termed as the “initial model” is stored on the central server. The participating devices (e.g., Node 1, as shown in Figure A1) contact the central sever to download a copy of the model and use local data to train the model. This helps to improve the model on the device using local data. After the model is trained, model’s parameters are sent back to the central server. The other devices (e.g., Node 2 and Node 3, as shown in Figure A1) can similarly pick up the global model and perform similar tasks of updating the model with their local data. The key advantage of this model is that the local data is not shared with the other devices, ensuring data privacy, and the model is also better through training on multiple data sets.

In summary based on the above steps, a global model is trained on data belonging to hundreds and thousands of patients, without them ever having to share data.

Centralized server (Aggregator)

The aggregation of all the models that are trained individually at respective edge locations takes place at the central server. A central server takes the responsibility to coordinating the training of all edge devices (nodes) that are part of the training process. The central server creates a global model by receiving model parameters from individuals edge devices (nodes) and consolidating all these models. In addition, the central server also defines the architecture of the global model. Some implementations deploy central servers on the cloud, while others use local data centers to deploy these servers. The centralized server maintains and shares the updated model. The information for each node remains private since only model parameters are being transferred between nodes. This is crucial for RHM use cases where protecting patient privacy comes first.

In summary, federated learning enables healthcare providers to train AI models on vast amounts of patient data while ensuring that sensitive information remains private and secure. This privacy-centric approach, supported by encryption and decentralized learning, is particularly valuable in applications like e-doctor, where protecting patient confidentiality is paramount.

Appendix 2: Heart attack database

This dataset focuses on heart disease diagnoses and contains essential information to help predict the presence of heart disease based on various clinical and demographic factors. The dataset provides valuable insights into the diagnosis and treatment of heart disease.

Dataset information

- Name: Heart Disease Prediction Dataset.
- Source: Collected for research purposes, typically focusing on the correlation between clinical factors and the presence of heart disease.
- Data Collection: The dataset includes clinical data points from various patients used to train machine learning models to predict heart disease.

Key features for heart disease diagnosis

The dataset includes multiple diagnostic variables that can help in detecting heart disease, including:

1. Age: The age of the patient.
2. Sex: Gender of the patient (male or female).
3. Chest Pain Type (cp): Type of chest pain experienced by the patient (e.g., typical angina, atypical angina).
4. Resting Blood Pressure (trestbps): The patient's resting blood pressure.
5. Cholesterol (chol): Serum cholesterol levels in mg/dl.
6. Fasting Blood Sugar (fbs): Whether the patient's fasting blood sugar is greater than 120 mg/dl.
7. Resting Electrocardiographic Results (restecg): Results from the resting electrocardiogram (normal or abnormal).
8. Maximum Heart Rate Achieved (thalach): The maximum heart rate achieved during a test.
9. Exercise-Induced Angina (exang): Whether the patient experienced angina during exercise.
10. Oldpeak: Depression induced by exercise relative to rest.
11. Slope of the Peak Exercise ST Segment (slope): The slope of the ST segment during peak exercise.

Key variables

- Target: The primary target variable indicating the presence of heart disease (1 for presence, 0 for absence).
- Ca: The number of major vessels colored by fluoroscopy.
- Thal: A blood disorder variable (normal, fixed defect, or reversible defect).

Use cases

This dataset is widely used for:

- Predictive modeling: Developing machine learning models to predict the presence of heart disease based on clinical features.
- Medical research: Understanding the correlation between different clinical features and heart disease outcomes.
- Healthcare decision support: Aiding healthcare professionals in identifying high-risk patients based on their clinical profiles.

This dataset offers rich clinical data for building models to aid in the early detection and diagnosis of heart disease, ultimately improving patient outcomes.

Appendix 3: Cancer prediction dataset

The Cancer Prediction Dataset is commonly used in medical research and machine learning applications for predicting whether a breast tumor is benign or malignant based on clinical features. The key features of the dataset are as follows.

Key columns

- mean_radius: The average radius (size) of the tumor.
- mean_texture: The mean texture, representing variance in the gray-scale values of the tumor.
- mean_perimeter: The average perimeter of the tumor.
- mean_area: The average area covered by the tumor.
- mean_smoothness: Measures the variation in radius lengths across the tumor.

Target variable

- diagnosis: The target variable, where 0 represents a benign tumor, and 1 indicates a malignant tumor.

Use cases

This dataset is widely used in medical research and machine learning tasks for early detection of breast cancer. It is ideal for building classification models to predict whether a tumor is malignant or benign based on clinical features. The dataset supports the creation of clinical decision support models, aiding in faster diagnosis and improved treatment outcomes for breast cancer patients.

Appendix 4: Deep venous thrombosis (DVT) dataset from the framingham heart study

This dataset focuses on Deep Venous Thrombosis (DVT) diagnoses collected as part of the Framingham Heart Study from 1994 to 2012. The dataset provides crucial insights into DVT diagnosis and treatment across various cohorts.

Dataset information

- Name: SOE Pulmonary Embolus and Deep Venous Thrombosis (DVT) Dataset.
- Accession: pht000744.v5.p12.
- Study: Part of the Framingham Heart Study, a landmark longitudinal study focused on cardiovascular health.
- Data Collection Period: 1994-2012.

Key diagnostic methods for DVT

The dataset includes data from multiple diagnostic methods used to detect and confirm DVT, including:

1. Venogram: An X-ray to detect blood clots in the deep veins of the legs.
2. Doppler Ultrasound: A non-invasive test using sound waves to detect blood flow and identify clots.
3. Impedance Plethysmography: A test measuring blood volume changes to detect clots.
4. Fibrinogen Leg Scan: A scan used to locate blood clots in the legs.

Key variables

- PE013 (Venogram: Result): Results of the venogram test indicating the presence of DVT.
- PE014 (Venogram: Location): The location of the blood clot identified by the venogram.
- PE016 (Doppler Ultrasound: Result): Results from the Doppler ultrasound, indicating evidence of DVT.
- PE017 (Doppler Ultrasound: Location): The location of the clot identified via Doppler ultrasound.
- PE019 (Impedance Plethysmography: Result): Results from impedance plethysmography testing for DVT.
- PE020 (Impedance Plethysmography: Location): The specific location of the clot identified through impedance plethysmography.
- PE022 (Fibrinogen Leg Scan: Result): Results from the fibrinogen leg scan used for DVT detection.
- PE023 (Fibrinogen Leg Scan: Location): The location of the clot detected by the fibrinogen leg scan.

Use cases

This dataset is valuable for research into the diagnosis and management of Deep Venous Thrombosis (DVT) in clinical settings. It allows researchers to:

- Assess the efficacy of different diagnostic methods for detecting DVT.
- Study the occurrence and outcomes of DVT over time in a longitudinal study.
- Explore demographic and clinical factors associated with a higher risk of DVT.

Appendix 5: Local data for better health

The 500 Cities dataset, released in 2018, provides small area estimates for 27 measures related to unhealthy behaviors, health outcomes, and the use of preventive services. It includes data from the 500 largest US cities and approximately 28,000 census tracts.

Key variables

- City Name: Name of the city included in the dataset.
- State Abbreviation: Abbreviation of the state where the city is located.
- Measure: The health measure being assessed (e.g., high blood pressure, smoking, mammography use).
- Data Value: The data value for the specific health measure, represented as a percentage or count.
- Geographic Level: Identifies whether the data is at the US, city, or census tract level.
- Data Source: Source of the data (e.g., Behavioral Risk Factor Surveillance System-BRFSS).

Measures of interest

- Health behaviours: Includes measures like binge drinking, smoking, and physical inactivity.
- Health outcomes: Tracks chronic conditions such as diabetes, stroke, cancer, and heart disease.
- Preventive services: Data on services like cholesterol screening, mammography, and dental checkups.

Use cases

This dataset is a key resource for:

- Public health research: Tracking chronic disease patterns and preventive behaviors.
- Geospatial analysis: Exploring variations in health outcomes across different regions.
- Predictive modeling: Identifying emerging health issues and risk factors in specific geographic areas.

The dataset was part of the CDC’s 500 Cities project, aimed at improving public health by providing localized data for targeted interventions.