MedBot – An NLP based ChatBot for Diabetes Prediction

Sreekar K
Amrita School of Computing
Amrita Vishwa Vidhyapeetham
Coimbatore, India
cb.sc.p2aie23004@cb.students.amrita.edu

Aswin V
Amrita School of Computing
Amrita Vishwa Vidhyapeetham
Coimbatore, India
cb.sc.p2aie23001@cb.students.amrita.edu

Vishnu Narayanan
Amrita School of Computing
Amrita Vishwa Vidhyapeetham
Coimbatore, India
cb.sc.p2aie23015@cb.students.amrita.edu

Dr. T Senthil Kumar
Amrita School of Computing
Amrita Vishwa Vidhyapeetham
Coimbatore, India
t_senthilkumar@cb.amrita.edu

***Abstract*— In this age of technology, with digital devices being accessible by almost everyone, applications of technology for many tasks have made out lives better. With the advent of Machine Learning and advanced algorithms at our tips, the applications are immensurate. Such an application can be found in the health-care sector. The need for better and accessible health-care is of immense precedence. In this paper, we employed Machine Learning algorithms like Logistic Regression and other state of the art techniques to predict diabetes, given patient data. We have employed PIMA Indian Diabetes Dataset which is furnished through an interactive chat-bot as they have emerged as a game-changer in the healthcare industry, streamlining appointment scheduling, providing interactive medical advice, and enhancing patient engagement, using Telegram API, resulting in an accuracy of 77%.**

Keywords*— *Chat-Bot, Logistic Regression, Natural Language Processing, Diabetes, Telegram, Health Care

I. INTRODUCTION

Diabetes can be characterized as a culmination of the problems that involve insulin, a crucial hormone produced by human body to regulate blood glucose levels. Michael Dansinger, MD says that [1] basically, insulin is developed by an organ that is present at the rearward of stomach, called Pancreas. It helps in utilizing the fats and sugars we get from food, by releasing insulin into the blood. However, the condition Diabetes occurs when one of the following occurs, if there is the Pancreas fail to produce insulin, or when it is unable to produce necessary amount of insulin, or sometimes if the body is inert to insulin (insulin resistance).

Diabetes is regarded to be incessant, something that is life-long. People diagnosed with diabetes have to personally take much care continuously to stay healthy. India alone has an estimated 77 million people suffering from diabetes [2], and that number being a staggering 366 million worldwide. Michael Dansinger, MD also elaborates in his article on how vital the role of insulin is. Whatever we consume, after being digested is converted to a simple sugar compound called "Glucose", which is transported to all the cells in the body via bloodstream. Insulin carefully regulates the amount of glucose in the blood. Pancreas release insulin quite often, albeit in small amounts. When the level of glucose in blood spikes, more insulin is produced to transfer more glucose to the cells. However, this would cause the blood-glucose levels to plummet. To avoid any further harm, which could be the result of a further drop in blood-glucose levels (a condition termed as hypoglycemia), we feel hungry, to eat more food or sometimes, the stored glucose from the liver is used. Thus, this condition of non-response to insulin is diabetes. Medically, it is referred to the situation where the blood-glucose levels would be 126 milligrams per deciliter (mg/dL) or higher, on an empty stomach for the previous night.

Diabetes is generally categorized into four types, based on the onset conditions and how the human body responds to it. When the blood-glucose levels are higher than but not as high to be classified as diabetes, it is termed as impaired glucose tolerance (prediabetes). This is usually asymptomatic, nevertheless this would most likely result in type-2 diabetes if not identified and diagnosed at the earliest. A specific type of cells, called beta-cells, present in pancreas are responsible for insulin production. But for some people there cells are completely absent or destroyed by the immune system as a result of some auto-immune condition. This condition is known as type-1 diabetes. People diagnosed with type-1 diabetes have to use insulin injections to regulate blood-glucose levels. Type-2 diabetes however, is a different condition. Here, the pancreas produce insulin, but it is either not being absorbed by the body or if the produced quantity is not enough, it is termed as type-2 diabetes. Gestational diabetes is found in women during pregnancy. The hormonal changes induced by pregnancy can alter the working of insulin in some women. Typically primary screening is done during pregnancy, but if left unidentified, it would transform into type-2 diabetes.

Type-1 Diabetes accounts for around 5-20% of the total, these people would be in the need of exogenous insulin at the earliest after preliminary diagnosis, for their survival, as stated by Marian Rewers [3] in his research. Type-2 diabetes however, is the most prevalent of all. It still is the major cause of complications induced by diabetes, such as non-traumatic amputations, blindness, and kidney failure. It was initially believed to only affect adults, but recently due to obesity in children, Type-2 diabetes is being observed in teenagers and chikdren as well. The concerning issue is that many people are not aware of the symptoms, which if not diagnosed early would result in some serious complications like heart stroke, neuropathy, reduced blood flow, along with many other issues. This brings us to the point that early detection of diabetes is crucial which could be decisive for further treatment and wellbeing.

The process starts by preparing the dataset, followed by data pre-processing such as handling missing values and imputing categorical values, combined with standardization. A number of tools have been used to perform feature selection. Finally the classification techniques have been used and their

performance is presented and assessed. This paper is organized into various sections, Section 2 being Motivation, Section 3 being Literature Review where we explore related works in the area and involved implications. This is followed by Methodology in Section 4, wherein we explore the dataset and the algorithms involved in detail. Finally Section 5 concludes the paper.

II. Motivation

To address the issue of early detection of diabetes, which could prove vital for immediate diagnosis, we employed advanced computing techniques that involve sophisticated mathematical concepts and pattern matching techniques. Machine Learning paradigms have revolutionized many disciplines where manual work would rather have been laborious and error prone. As we can see in the work of C. J. Harrison and C. J. Sidey-Gibbons [4], how machine learning algorithms have influenced the medical diagnosis landscape. They have made disease prediction much faster and reliable, which could potentially save many lives. We have employed a number of machine learning techniques to detect diabetes based on the user input. The user is presented with a number of parameters which they have to input to the algorithm, which would then classify if the user is diabetic or not. This would have a major advantage in early screening as the process would be agile and streamlined. If detected as positive, the user can immediately consult a certified healthcare professional for treatment.

Also, to facilitate enhanced portability and sustained interactive facilities for the user, we incorporated the prediction algorithm within a chat-bot, based on Natural Language Processing techniques (NLP). There have been quite a number of chat-bots that have witnessed diverse applications in numerous fields that drew the attention of many people. They are successful for being interactive and reliable while also being portable, making it easy to deploy and use it without much hardware dependencies. A number of Machine Learning algorithms have been tried and tested, such as Logistic Regression, Decision Trees, K-means clustering with KNN, etc. Our goals have been towards reliable accuracy whilst having a light-weight model. This would facilitate easier deployment on a variety of devices, without the need of powerful hardware. In this paper, we have employed PIMA Indians Diabetes Dataset [5]. It is a widely recognized dataset for diabetes analysis. This dataset has been pre-processed and worked over with multiple machine learning algorithms which was eventually integrated with the popular internet messaging service Telegram [6], which provides a native support for chat-bots within the application.

III. Literature Review

A. ML Techniques for Diabetes Classification

Diabetes classification, just like many other medical classifications involve quite a lot of skilled human labor for initial screening. It basically involves checking blood-glucose levels through blood tests. Typically, two tests are done, one with fasting and one after taking food, as explained by Deborah J Wexler, MD, MSc [7]. This process will naturally take time for the results to be processed by medical personnel and given to the user. That is where ML techniques come into play. Conrad J. Harrison [4] has presented meticulously in his review, how ML techniques like NLP can be used to extract insightful details from rather unstructured data. This technique would let the unstructured data be transformed into meaningful datasets that can be worked with various other standard ML algorithms.

Similar approach is evident in the work of I. Zafar et al. [8] where his team from Virtual University of Pakistan had investigated the use of Deep Learning (DL) in genomics and biomedicine. This threw light on how DL frameworks like Keras, PyTorch, Tensorflow can be used to extract patterns from humongous datasets that will have been generated since many years of clinical study. These patterns would help many researchers. His study primarily focuses on using Deep Neural Networks (DNNs) for illness detection, whilst also reviewing a number of deep learning techniques, optimization approaches and architectures. His study has explicitly emphasized the use of Explainable AI (XAI) to enhance medical personnel's understanding of DNNs so they can fine-tune them based on their experiential knowledge. His text includes a wide array of datasets, plant species, animal breeds, and human features, to name a few. Even though the efficacy of DL models can be acknowledged, the researchers insist to improve and alter these approaches to be used in agriculture and customized medicine.

Another interesting research was done by H. Naz and S. Ahuja [9], where they have conducted a series of tests using a number of ML and DL techniques, like ANNs, Decision Trees, Naïve Bayes and DL networks. The author was motivated by the serious complications caused by diabetes when left lingering without identifying it at the earliest. They have worked on the reputed PIMA Indian Diabetes dataset, garnering a number of useful insights. They have achieved an accuracy of 98% using the Deep Neural network. The importance of omnics data for further reliability has been stressed upon in this study.

Likewise, in the works of M. Maniruzzaman. et al, [10] from Khulna University, we have used ML techniques like random forest classifiers and logistic regression to predict diabetes. Using the method of p-value and ratio of odds, they were able to fine-tune the existing framework. This study involved an independent dataset curated by collecting samples from around 6500 participants (657 who were diabetic and rest were controls). This study highlights the significance of precise risk factor identification and offers insightful information on diabetes prediction using mathematical models.

Another extensive study performed by V. Chang, et al, [11] advocates the elaborate advantage of employing PIMA Indian Diabetes dataset. They have done a meticulous analysis on the dataset and have proposed a system based on the Internet of Medical Things (IoMT) environment. They have delineated three particular ML models, J48 Decision Tree classifier, Random Forest classifier, Naïve Bayes classifier. It has been totally trained and tested on the PIMA Indian Diabetes dataset, programmed in R language. Their study can be concluded with the Naïve Bayes performing well with a more selective number of features for two-class classification, while Random Forest working better when involving multiple features. The training sample consisted 538 samples whereas test set has 230 samples. Naïve Bayes model has achieved an accuracy of 77% which was higher than the analyzed models in this research work.

The work led by B. C. Han, et al [12] at Seoul National University, used topic modeling and Machine Learning techniques to predict the outcomes and course of Diabetes

Table 1 A concise summary along with potential research gaps

Ref No.	Dataset	Algorithm	Research Gap
[3]	Manually curated clinical and ethnographic data pertaining to population in Africa and US	A thorough review is presented analyzing various attributes that pose a challenge in diagnosing Type 1 diabetes.	The paper's research gap is the need for a thorough analysis and synthesis of the available data to enhance Type 1 Diabetes (T1D) diagnosis across a range of ethnic groups. This analysis should concentrate on resolving difficulties in interpreting commonly used diagnostic results and highlighting the significance of early identification to reduce the morbidity associated with postponed insulin treatment.
[4]	Drug Review Dataset from the University of California	Latent Dirichlet Allocation (LDA), Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression	The paper's research gap is the need for more investigation and improvement of natural language processing (NLP) techniques applied to unstructured medical text data, particularly in enhancing the efficacy of supervised machine learning algorithms for sentiment analysis of drug reviews and tackling the difficulties involved in predicting drug ratings in a variety of datasets gathered from various sources. This will help to advance the field's practical and repeatable methodologies.
[9]	PIMA Indian Diabetes Dataset	ANNs, Decision Trees, Naïve Bayes and DL Networks.	To improve the precision and efficacy of early detection and prognosis tools, the research gap in the report focuses on the necessity of more research into the integration of varied data types, such as omics data, in machine learning algorithms for diabetes prediction.
[10]	Diabetes data from 2009–2012 derived from the National Health and Nutrition Examination Survey (NHANES)	Logistic regression (LR), Naïve Bayes (NB) , Decision Tree (DT), AdaBoost (AB), and Random Forest (RF)	The paper's research gap is the need for additional investigation and improvement of machine learning-based systems for diabetes prediction, with an emphasis on finding the best combinations of classifiers and feature selection strategies as well as resolving issues related to risk factor identification in order to improve the predictive models' overall applicability and accuracy.
[11]	PIMA Indian Diabetes Dataset	Naive Bayes classifier, random forest classifier, and J48 decision tree models	By highlighting the significance of interpretability and offering a comparative analysis of interpretable supervised ML models, the paper fills a research gap related to addressing the trust issues surrounding machine learning applications in healthcare, specifically in the context of diagnosing diabetes. However, it also leaves room for additional investigation and decision process refinement to improve overall model performance and acceptance within healthcare sectors.
[12]	Around 174,000 clinical notes were collected from Seoul National University Hospital outpatient clinic.	Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XG-Boost).	The study's research gap is the need for more investigation and development of machine learning models, particularly topic modeling like latent Dirichlet allocation (LDA), to forecast the development of complications related to diabetes mellitus (DM). This research should be focused on enhancing predictive performance and generalizability across different kinds of complications.
[13]	Kaggle EyePACS dataset with 35,126 retinal pictures has been used.	A new SE-ResCA-GTNet model is proposed and Gazelle Optimization (GO) algorithm is used to fine-tune the proposed classifier	The study's research gap is in the need for improved techniques and strategies for the early detection and classification of diabetic retinopathy (DR). These techniques should integrate transformer networks, precisely segment important features, and apply the Gazelle Optimization algorithm in order to outperform current state-of-the-art techniques and increase accuracy in diagnosing the disease's various severity levels.
[14]	A custom dataset has been garnerd from Reddit. The dataset contain comments of January, year 2015. The format of data is in JSON format.	BRNN with Attention model	The study fails to specifically identify and discuss the difficulties or constraints associated with using Bidirectional Recurrent Neural Networks (BRNN) with attention layers for an Assistant Conversational Agent. It also fails to address possible areas for future research and development related to Chat-Bot performance and English-to-English translation.

[15]	A survey based on electronic health records.	This paper explores all the existing methods used in this area	This paper's research gap is the need for better deep learning methods development and applications in the healthcare industry. It focuses on developing comprehensive and meaningful interpretable architectures to improve the bridge between human interpretability and deep learning models, and it addresses issues related to ease of understanding for domain experts and citizen scientists.

Mellitus (DM). This research has fostered a decent degree of predictive performance with emphasis over non-alcoholic fatty liver disease, diabetic retinopathy, and diabetic nephropathy. Around 174,000 clinical notes were collected from Seoul National University Hospital outpatient clinic. These notes were scanned and analyzed by the electronic medical record (EMR) system. An extensive degree of pre-processing was applied over the dataset, as most of the notes were written in Korean syntax, also the names of the drugs were different even-though they refer to the same generic chemical. Algorithms such as Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XG-Boost) have been implemented. By transferring topic models from training to test data, the study effectively predicts the probability of complications by using training data to identify similarities between topic structures and future problems. The authors press on the the promise of this method for improving prognosis from clinical notes and call for more research to understand the elements impacting forecasts.

S. Karthika, et al [13] has presented an improved attentional global transformer network with ResNet_101 for automated diagnosis of diabetic retinopathy. In order to prevent blindness, the study discusses the vital early identification of diabetic retinopathy, a frequent eye condition associated with diabetes. The dataset used was acquired from Kaggle, which is EyePACS dataset that has 35,126 retinal pictures. This proposed system uses a Squeeze Excited ResNet_101 architecture with a Crossfield Attention Global Transformer Network, using many models like DenseNet121 and ResNet50. This model presents respectable accuracy in determining the severity of the illness after testing on 40 real-time fundus imaging datasets. Despite having these caveats, longer training durations and large dataset requirements, the team suggests for further work on dataset augmentation, domain adaption, and integrating the model into a real-time smartphone app for clinical DR eye exams.

B. Chat-Bot for medical applications

Chat-Bots are tightly programmed structures that actively respond to user queries by presenting information, mostly in text format [6]. R. Agarwal and M. Wadhwa [16] have explored a number of approaches for chat-bot development, including rule-based and neural network based methodologies. They have explored the changes in chat-bots from primitive to advanced level of chat-bots. It draws attention to evaluation criteria like confusion and bi-lingual evaluation understudy. This article represents a stark contrast retrieval-based and generative approaches, highlighting the chat-bot's primary responsibility of providing pertinent answers. For conversational modeling, the neural network approach is explored, which allows for context-sensitive answers. The last section of this study highlights recent developments in the subject and makes some recommendations for future development. It encourages further investigation into a range of conversation modeling facets, thereby broadening the research community's comprehension of chat-bot development.

Another study by S. Kusal.et al, [17] from Symbiosis International University proposed an elaborate analysis of conversational agents based on Artificial Intelligence (AI). The collection of Machine Learning, Deep Learning, and Natural Language Processing had helped Conversational Artificial Intelligence (AI) to change the face of Human-Computer Interaction (HCI). They propose that the agents must employ sophisticated natural language processing and machine learning to understand user emotions and context in order to produce custom tailored responses and have conversations that can replicate human conversations. This review covers a number of implementation strategies, activities, and the possible use of emotions to improve user experiences. Based on an analysis of 5000 postings, the authors have also addressed the use of publicly available datasets and deep learning models, pointing out research gaps and making recommendations for future paths.

IV. METHODOLOGY

A. Description Of The Dataset

We have employed PIMA Indian Diabetes Dataset [5]. Despite having access to many other datasets, PIMA Indian Diabetes Dataset has been popular for benchmarking diabetes classification models as this is considered to be a significant representative of the global health. PIMA Indians refer to the Native American group present in Mexico and Arizona, USA. They are incidentally found with high rates of Diabetes Mellitus. It consists of data corresponding to PIMA Indian females older than 21 years. The dataset has been downloaded from Kaggle (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database). This is licensed under CC0, a public domain license. The data values corresponding to features are completely anonymous, making it a concrete dataset focusing on user's privacy. It has 8 independent feature variables and one dependent variable vector which corresponds to the result of the patient having diabetes or not. A total of 768 patient's data is furnished of which around 268 are positively tested with diabetes and the rest are non-diabetic. A generic view of the dataset is given in Figure 1 and Table 1.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1 PIMA Indian Diabetes dataset

B. Data-Preprocessing

This dataset is solely comprised of numeric data, spread over 768 rows and 8 columns. When we plot a box-whisker plot, as in Figure 2, we can see that there are some outliers that

Table 2 Description of the features

Feature	Description	Data Type	Range
Pregnancies	Number of times the patient was pregnant.	Integer (64)	0 – 17
Glucose	Blood-Glucose concentration in an oral glucose tolerance test (GTIT) for 2 hours.	Integer (64)	0 – 199
BloodPressure	Blood Pressure (mm/Hg).	Integer (64)	0 – 122
SkinThickness	Skin fold thickness measured at the triceps muscle (mm).	Integer (64)	0 – 99
Insulin	Level of insulin in 2 hours during the Glucose test.	Integer (64)	0 – 846
BMI	Weight of the person in kg / Height of the person in meters.	Float (64)	0 – 67.1
DiabetesPedigreeFunction	The likelihood of a person getting diabetes based on his/her family history.	Float (64)	0.078 – 2.42
Age	Age in years.	Integer (64)	21 – 81
Outcome	A binary value representing if the person tested is diabetic positive or not.	Integer (64)	0 – 1

need to be handled, which generally would be zero-values or missing values. This dataset has no records having missing or null values, however it has some inconsistent values corresponding to features such as 'Insulin', 'SkinThickness', 'Glucose', and 'BloodPressure'. It has some zero-values instances which are practically erroneous. To mitigate the impact of these missing values, the dataset is meticulously analyzed, and appropriate strategies are implemented.

In the preprocessing pipeline, instances where 'Insulin' levels are zero-values are identified and replaced with the mean value corresponding to the 'Insulin' level calculated from non-zero instances. This approach ensures that missing 'Insulin' values are imputed with representative values, enhancing the dataset's overall integrity. Similarly, for attributes like 'SkinThickness', 'Glucose', and 'BloodPressure', zero-valued entries are indicative of missing or invalid data. To rectify this, zero-valued entries are substituted with the mean values computed from non-zero instances of the respective attributes. By imputing the erroneous missing values with estimates calculated from the mean of the feature vectors would make the dataset consistent and robust.

Likewise, some features like as 'BMI', 'SkinThickness', and 'Age' may not significantly contribute to the predictive task at hand or may introduce unnecessary noise into the modeling process. As a result, these attributes are strategically removed from the dataset using the 'drop' function in the pandas library. This feature selection step streamlines the dataset, focusing on attributes that are most pertinent to the diabetes prediction task. The dataset after imputation can be seen in Table 2.

C. Classifiers

The following text gives a detailed description of the classifiers that were applied to the chosen features.

- Logistic Regression

  Logistic Regression is a fundamental statistical technique used for binary classification tasks, where the goal is to predict the probability that a given input belongs to one of two possible classes. It's widely used in various fields including healthcare, finance, and marketing due to its simplicity and interpretability. The core idea behind logistic regression is to model the relationship between the independent variables (features) and the probability of a

Table 3 Description of data in PIMA Indian Diabetes dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	3.845052	120.8945	69.10547	20.53646	79.79948	31.99258	0.471876	33.24089	0.348958
std	3.369578	31.97262	19.35581	15.95222	115.244	7.88416	0.331329	11.76023	0.476951
min	0	0	0	0	0	0	0.078	21	0
25%	1	99	62	0	0	27.3	0.24375	24	0
50%	3	117	72	23	30.5	32	0.3725	29	0
75%	6	140.25	80	32	127.25	36.6	0.62625	41	1
max	17	199	122	99	846	67.1	2.42	81	1

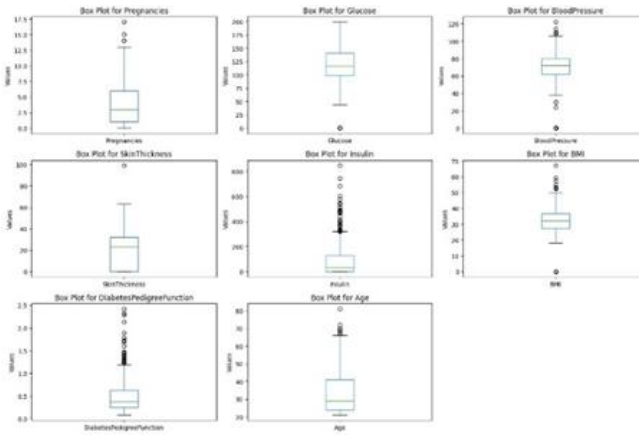

Figure 2 Box-Whisker plot describing the feature spread

binary outcome using a logistic function, also known as the sigmoid function, as shown in Equation 1, where e represents the Euler constant and x represents input feature vector.

$$S(x) = \frac{1}{1+ e^{-x}} \quad (1)$$

Logistic Regression models the relationship between the independent variables (features) and the dependent variable (binary outcome) using a logistic function. The logistic regression hypothesis is defined as a function that maps a real-valued number to a number between 0 to 1.

Central to logistic regression is the sigmoid function, which transforms the output of the linear combination of features into a probability score bounded between 0 and 1. The sigmoid curve's characteristic "S" shape ensures the interpretability of predicted probabilities and facilitates the delineation of decision boundaries in binary classification tasks. The sigmoid curve, shown in Figure 5 plays a pivotal role in logistic regression by conferring the capability to model non-linear relationships between features and probabilities. Its ability to smoothly converge to 0 or 1 as the input varies ensures the generation of well-calibrated probability estimates, essential for making informed decisions in binary classification scenarios.

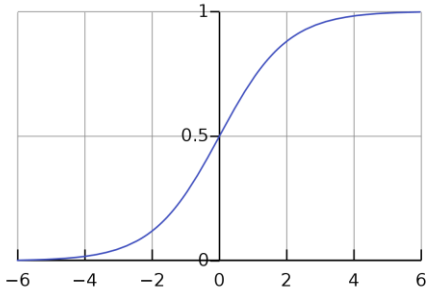

Figure 3 Sigmoid Curve

Logistic regression models are trained using optimization techniques to minimize the cost function, thereby learning optimal parameter values. Model evaluation entails assessing performance metrics such as accuracy, precision, recall, and F1 score to gauge the model's effectiveness in classifying instances. There have been numerous applications where logistic regression has been employed for medical related classifiers, as we can see in the work of C. Zhu.et al [18], wherein they employed Logistic Regression for diabetes classification for a custom dataset.

- Decision Tree

Decision Trees represent a fundamental technique in the realm of machine learning and data mining, offering an intuitive and interpretable approach to classification and regression tasks. Renowned for their simplicity and transparency, decision trees find widespread application across various domains, including healthcare, finance, and marketing. They have an elaborate application in Medical field as illustrated by Rodionov, Andrei and Mukhitdinova [19].

A Decision Tree is a hierarchical structure composed of nodes that represent decision points based on feature attributes. Through a series of binary decisions, the tree partitions the feature space, ultimately leading to the assignment of instances to specific classes or predicting continuous target variables. At each decision node, the decision tree algorithm selects an optimal split criterion based on impurity measures such as Gini impurity or entropy. This criterion determines the most informative attribute and value to split the data, thereby maximizing class purity within each resulting subset. Tree Pruning Techniques: To prevent overfitting and enhance model generalization, decision trees incorporate pruning techniques such as cost-complexity pruning and reduced error pruning. These techniques iteratively remove nodes with minimal contribution to overall model performance, resulting in simpler and more interpretable trees.

Model Interpretability: One of the key advantages of decision trees lies in their inherent interpretability. The hierarchical structure of decision trees enables straightforward interpretation of decision paths and feature importance, facilitating insights into the underlying data patterns and decision-making process.

Decision trees find diverse applications across a spectrum of domains, ranging from medical diagnosis, as done by V. Chang, et al [11] and risk assessment to customer segmentation and fraud detection. Their ability to handle both categorical and numerical data, along with their interpretability, makes decision trees a preferred choice for complex decision-making tasks.

In recent years, advancements in decision tree algorithms, such as ensemble methods like Random Forests and Gradient Boosting Machines, have further enhanced the predictive performance and robustness of decision tree models. Future research directions may explore novel techniques for handling imbalanced data, integrating decision trees with deep learning architectures, and enhancing model interpretability.

- K-Means with Convolutional Neural Network

K-Means clustering and Convolutional Neural Networks (CNNs) serve separate functions in machine learning applications. In unsupervised learning, K-Means clustering serves as an unsupervised clustering algorithm, helping to divide a dataset into K numbered clusters based on feature space similarities (generally distance based). Its uses generally involve image compression and

segmentation. CNNs, on the other hand, are a type of Deep Neural Network learning model that is optimized for grid-structured data, such as images and video data where the data can be partitioned into pixel-level elements. CNNs use convolutional layers to automatically learn feature hierarchies. While K-Means and CNNs normally work independently, there is significant synergy between them which we tried to explore. K-Means clustering can be used to initialize the parameters of a CNN, especially during unsupervised pre-training. K-Means are used to set initial filters in CNN's early layers.

Such an approach improves convergence during future supervised training, possibly leading to better performance in image classification tasks. The combination of K-Means and CNNs demonstrates how varied approaches may be used together to solve complicated problems in machine learning and computer vision.

- K-Means with K Nearest Neighbors

The relationship between K-means Clustering and K-nearest neighbors (KNN) algorithms has been interesting. Using 'K' as a guiding parameter, K-means Clustering efficiently divides data into clusters, exposing hidden patterns, which can be leveraged for medical data sets as referred by Cabello et al. [20] Consequently, the supervised algorithm KNN portrays its prediction capability by averaging the results of its K nearest neighbors or taking into account the majority class.

When integrated together, K-means Clustering and KNN provide a robust technique for classification as proposed by Shaikh et al. [21]. K-means Clustering creates a foundation for meaningful data grouping, enabling KNN to carry out classification inside each cluster. Together, they not only deliver a decent accuracy but also help in presenting meaningful insights from a variety of datasets.

- Fuzzy C-Means with KNN

In my investigation of machine learning approaches, I frequently investigate the interaction between Fuzzy C-Means (FCM) clustering and K-nearest neighbors (KNN) algorithms as investigated by Mabel Rani et al. [22]. FCM, a powerful unsupervised learning method, excels at dividing data into fuzzy clusters, which allows for nuanced memberships. This contrasts with the conventional sharp borders of K-means. Meanwhile, KNN, a mainstay of supervised learning, refines predictions by taking into account the consensus of nearby instances.

The convergence of FCM and KNN creates a dynamic fusion, with FCM delineating fuzzy clusters and KNN navigating with precision in classification tasks within each cluster. This approach not only ensures accuracy in predictions, but also provides a nuanced perspective on uncovering patterns and insights from complex datasets.

D. Chat-Bot using Natural Language Processing Techniques

We have identified that deploying the model as a chat-bot would make it widely accessible and portable. Chat-Bots have become popular recently as efficient agents for human-computer interaction. The goal of this project is to create a smart chat-bot that can communicate with humans and also take input parameters for diabetes classification. Chat-Bots have seen numerous applications in many disciplines, as seen in the research carried by R. Pradeep.et al, [14]. A Natural Language Processing (NLP) technique based chat-bot would need data to extract features and form tokens which would be used to dynamically generate strings based on user queries. We have employed web-scraping techniques using Python3 scripts and have compiled a collection of text curated from Wikipedia.com [23]. This is a public encyclopedia whose authenticity is concurred by many. The data scraped from the website is saved in a text file, which would provide the user necessary information regarding diabetes when questioned upon.

Based on user input, the given code creates a functional chat-bot that communicates with people. The required libraries are first imported by the code, which includes scikit-learn modules for vectorization, similarity computations and NLTK for tasks involving natural language processing. It reads the text from the text file generated by web-scraper and lowercases it to avoid inconsistencies which is then tokenized it into words and phrases. To lemmatize words, the NLTK WordNetLemmatizer() is started. By lemmatizing the tokens, the LemTokens() method makes sure that variants of the same term are handled consistently.

By changing the text to lowercase, eliminating punctuation, and lemmatizing the words, the LemNormalize() method normalizes the text. The greeting() method determines whether any preset words or phrases are included in the user's input. The chat-bot responds with a random greeting from the pre-programmed list of options if a greeting is recognized. After processing the user's message, the answer() method uses cosine similarity and TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to provide a response.

We have employed a cloud based system that has a server which is integrated with Telegram’s API. The user can interact with the chat-bot agent which stores the client data upon input. Then the ML module is invoked in the server which runs the model. The result is sent via API to the Telegram client of the user. Figure 4 depicts the basic architecture of the system. Figure 5 gives an outline of the working of the system.

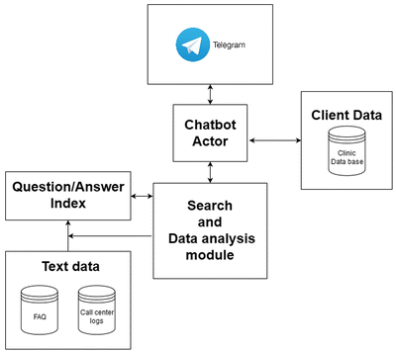

Figure 4 Architecture of the Telegram Chat-Bot

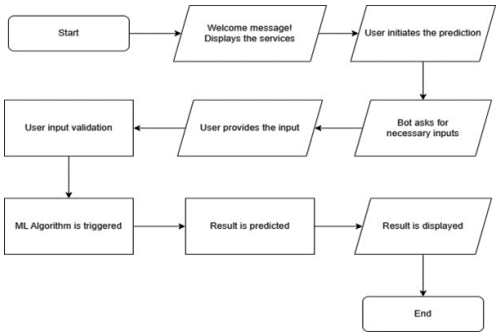

Figure 5 Model Diagram of the system

It computes the TF-IDF vectors for each sentence and appends the user's response to the list of sentence tokens. The sentence that most closely resembles the user's input is determined via cosine similarity, and a response is produced based on that similarity score. User interaction is handled via the output_is_here() method. After processing the user's input and ensuring that no goodbyes or greeting messages are sent, it calls the relevant response routines. Should the user's input not correspond with any pre-programmed salutations or farewells, a response is generated by calling the response() method.

V. Results And Discussion

We elucidate the relative performances of the applied Machine Learning and Deep Learning models by comparing them using various standard metrics. We have considered some key metrics like accuracy and precision, but also focused on the ease of deployment of the models. As the application needs to be portable, the model would have a trade-off between portability and model complexity.

Table 4 Accuracy and Precision of various ML/DL models

Model	Split-Ratio	Accuracy	Precision
Logistic Regression	80:20	0.76	0.67
Decision Trees	80:20	0.72	0.70
K-Means Clustering with CNN	80:20	0.72	0.71
K-Means Clustering with KNN	70:30	0.99	0.99
Fuzzy C-Means with KNN	80:20	0.76	0.74

Accuracy of a model refers to its ability of identify all the samples correctly belonging to their respective classes. It is the ratio of the sum of true positives and true negatives to the total number of predictions made, as in Equation 2.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Whereas, precision refers to the percentage of all samples that have been correctly identified as true out of all those which were predicted as true, even if they belong to the false class.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

When we constructively compare the performances of the chosen models in Table 2, we identify that most of the chosen algorithms have given a decent accuracy within the range of 70% - 77%, with K-Means Clustering with KNN giving 99%, which points out to overfitting. We have thus discarded that model not only because of its unrealistic accuracy but also the inherent complexity involved with the Neural Network which makes it difficult for deployment over a chat-bot environment.

Algorithms like Logistic Regression and Fuzzy C-Means with KNN have performed almost similarly with an accuracy of 76%. We have deployed Logistic Regression for the chat-bot integration as it involves simple mathematical functions that can be run over any light-weight server.

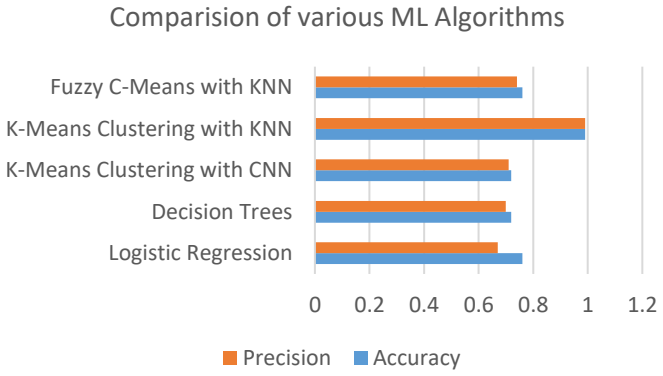

Figure 6 Performance of various models

We have chosen Logistic Regression in lieu of its light-weight nature and simple classifying function. As we have numerical attributes, it would be better to analyze and handle the underlying patterns and give us a binary classifier. Also, we have deployed the model over a server that is connected with a chat-bot running using Telegram API implemented using NLTK framework.

The chat-bot can be initiated by inputting the start message (/start). This would present the user with a Welcome Text. The user can choose to either start diabetes classification or get information regarding diabetes by choosing one of the two

options - /Diabetes (or) /Query for prediction and information respectively. When chosen to predict diabetes, the chat-bot asks the user to enter a string of values namely number of pregnancies, insulin, glucose, blood pressure, and diabetes pedigree function, as shown in Figure 5.

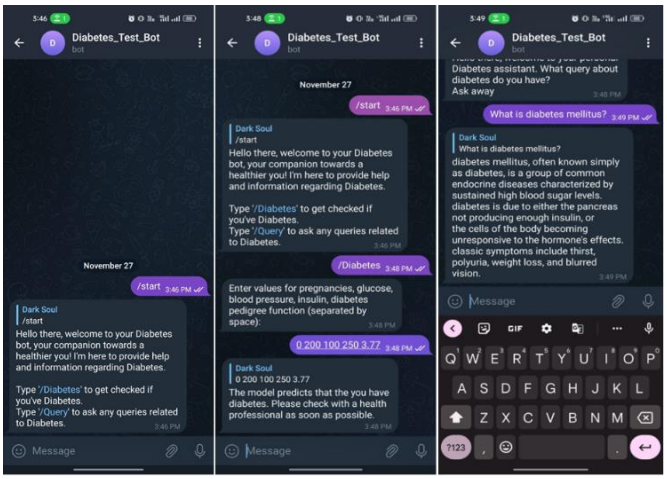

Figure 7 Screenshots of the Telegram Chat-Bot application

When the query action is initiated, the chat-bot parses the user input and uses NLTK framework to extract the tokens that would best address the user query with maximum confidence.

VI. Conclusion And Future Scope

We have chosen Logistic Regression in lieu of its light-weight nature and simple classifying function. As we have numerical attributes, it would be better to analyze and handle the underlying patterns and give us a binary classifier. Also, we have deployed the model over a server that is connected with a chat-bot running using Telegram API implemented using NLTK framework.

There can be a myriad applications of medical chat-bots that can streamline the medical industry by providing quick and reliable immediate attention-based care which could prove vital to save many lives. This paper presents an opportunity to implement an integrated system that can not only predict diabetes, but also many other diseases through an easy-to-use medium such as chat-bot. Advanced web-scraping techniques can be automated to fetch the data in real-time as this can be advantageous when the research in that area is ongoing and prone to changes. Advanced Machine Learning and Deep Learning techniques can be employed to classify diseases. This can be scaled upon to an extent to function as a fully-fledged personal first-responding health care manager.

VII. Appendix

Explainable AI (XAI) can significantly enhance the utilization of machine learning (ML) and deep learning (DL) techniques in the health care sector, as in the work of S S Band et al. [24], we can employ it for predicting diabetes, especially when implemented within a system such as a chat-bot that utilizes logistic regression. XAI is very important in the context of a proposed Logistic Regression-based Natural Language Processing (NLP) chat-bot to forecast diabetes, since it can greatly improve the system's efficacy and user acceptability. Here's how XAI may be used to enhance user engagement and the chat-bot's prediction abilities.

The chat-bot can provide clear and understandable forecasts, explaining to customers the reasoning behind risk evaluations based on variables such as age, body mass index, and family history. Furthermore, XAI helps determine which characteristics have the most influence on diabetes risk, enabling users to rank health indicators for follow-up and treatment. Through the identification of biases and mistakes, XAI guarantees accuracy and fairness across a range of user demographics, as postulated by Amman et al. [25], promoting confidence and openness in the chat-bot's decision-making process. Additionally, by including users in the prediction process and offering insightful explanations, XAI encourages user involvement and increases trust and adherence to the chat-bot's suggestions. XAI makes it easier to continuously develop models by integrating fresh insights and user feedback, which eventually improves prediction accuracy. By offering open and responsible AI systems, protecting user privacy, and encouraging ethical AI activities, XAI guarantees regulatory compliance and ethical concerns, as stated by Chadded et al. [26].

VIII. Acknowledgement

The authors would like to cordially thank Amrita Viswa Vidyapeetham's faculty and the research lab – Multicore Ware Academia Global Innovation Centre (MAGIC) at Amrita School of Computing, Coimbatore campus, Amrita Vishwa Vidyapeetham University India for providing the necessary infrastructure for carrying out the work.

IX. Data-Set Availability

The Dataset has been fostered from Kaggle.com [5] while the codes have been enclosed in Google Collaboratory, and also in our associated GitHub page [27].

References

[1] M. Michael Dansinger, "Diabetes Guide," 18 March 2023. [Online]. Available: https://www.webmd.com/diabetes/diabetes-basics.

[2] WHO, "Diabetes in India," WHO, 2023. [Online]. Available: https://www.who.int/india/health-topics/mobile-technology-for-preventing-ncds.

[3] Rewers Marian, "Challenges in Diagnosing Type 1 Diabetes in Different Populations," *dmj*, vol. 36, pp. 90-97, 2012.

[4] C. J. Harrison and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to natural language processing," *BMC Med. Res. Methodol.*, vol. 21, p. 158, 2021.

[5] U. M. LEARNING, "Pima Indians Diabetes Database," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[6] Telegram, "Bots: An introduction for developers," Telegram, [Online]. Available: https://core.telegram.org/bots.

[7] M. M. Deborah J Wexler, "Patient education: Type 2 diabetes: Treatment (Beyond the Basics)," uptodate.com, [Online]. Available: https://www.uptodate.com/contents/type-2-diabetes-treatment-beyond-the-basics/print.

[8] I. Zafar, S. Anwar, F. Kanwal, W. Yousaf, F. Un Nisa, T. Kausar, Q. ul Ain, A. Unar, M. A. Kamal, S. Rashid, K. A. Khan and R. Sharma, "Reviewing methods of deep learning for intelligent healthcare systems in genomics and biomedicine," *Biomed. Signal Process. Control*, vol. 86, p. 105263, 2023.

[9] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metab. Disord.*, vol. 19, p. 391–403, 2020.

[10] M. Maniruzzaman, M. J. Rahman, B. Ahammed and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Inf. Sci. Syst.*, vol. 8, p. 7, 2020.

[11] V. Chang, J. Bailey, Q. A. Xu and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, p. 1–17, 2022.

[12] B. C. Han, J. Kim and J. Choi, "Prediction of complications in diabetes mellitus using machine learning models with transplanted topic model features," *Biomed. Eng. Lett.*, vol. 14, p. 163–171, 2024.

[13] S. Karthika and M. Durgadevi, "Improved ResNet_{1}{0}{1} assisted attentional global transformer network for automated detection and classification of diabetic retinopathy disease," *Biomed. Signal Process. Control*, vol. 88, p. 105674, 2024.

[14] M. Dhyani and R. Kumar, "An intelligent Chatbot using deep learning with Bidirectional RNN and attention model," *Mater. Today*, vol. 34, p. 817–824, 2021.

[15] R. Miotto, F. Wang, S. Wang, X. Jiang and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, p. 1236–1246, 2018.

[16] R. Agarwal and M. Wadhwa, "Review of state-of-the-art design techniques for chatbots," *SN Comput. Sci.*, vol. 1, 2020.

[17] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, S. Mishra and A. Abraham, "AI-based conversational agents: A scoping review from technologies to future directions," *IEEE Access*, vol. 10, p. 92337–92356, 2022.

[18] C. Zhu, C. U. Idemudia and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Inform. Med. Unlocked*, vol. 17, p. 100179, 2019.

[19] A. Rodionov and M. Mukhitdinova, "The role of a decision tree as a method of artificial intelligence for the analysis of big data problems," *Economics and Innovative Technologies*, vol. 11, p. 400–407, 2023.

[20] S. Cabello and P. Giannopoulos, "On k-means for segments and polylines," 2023.

[21] M. H. Shaikh, K. J. Ho and F. Mustafa, "K-nearest neighbor based association data mining in healthcare correlated data systems," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 2022.

[22] A. J. Mabel Rani, C. Srivenkateswaran, M. Rajasekar and M. Arun, "Fuzzy C-means clustering on rainfall flow optimization technique for medical data," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 12, p. 180, 2023.

[23] "Diabetes," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Diabetes.

[24] S. S Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos and H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Inform. Med. Unlocked*, vol. 40, p. 101286, 2023.

[25] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai and P. consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, p. 310, 2020.

[26] A. Chaddad, J. Peng, J. Xu and A. Bouridane, "Survey of explainable AI techniques in healthcare," *Sensors (Basel)*, vol. 23, 2023.

[27] A. V. V. N. Sreekar K, "MedBot-Amrita," GitHub, [Online]. Available: https://github.com/sreekark99/MedBot_Amrita.

[28] R. Pradeep, S. Praveen Kumar, S. Sasikumar, P. Valarmathie and P. V. Gopirajan, "Artificial intelligence-based automation system for health care applications: Medbot," in *Advances in Intelligent Systems and Computing*, Singapore, Springer Singapore, 2022, p. 191–203.

[29] P. Patel and A. Macerollo, "Diabetes mellitus: diagnosis and screening," *Am. Fam. Physician*, vol. 81, p. 863–870, 2010.

[30] L. Jovanovic, "Screening and diagnosis of gestational diabetes mellitus," *Curr. Res. Diabetes Obes. J.*, vol. 4, 2017.

[31] M. Abedini, A. Bijari and T. Banirostam, "Classification of Pima Indian diabetes dataset using ensemble of decision tree, logistic regression and neural network," *Nternational J. Adv. Res. Comput. Commun. Eng.*, vol. 9, p. 1–4, 2020.

[32] "Retracted: Processing decision tree data using internet of things (IoT) and artificial intelligence technologies with special reference to medical application," *Biomed Res. Int.*, vol. 2023, p. 1–1, 2023.

About The Authors

Sreekar K is currently pursuing his Master of Technology degree in Artificial Intelligence, from Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India. He had finished his Bachelors from Jawaharlal Nehru Technological University, Hyderabad, Telangana, India fostering a Bachelor of Technology in Computer Science and Engineering. He is passionate about Algorithm Design and Machine Learning concepts. He had worked in a number of companies, both as an intern and as a full-time employee, garnering a wide array of technical prowess.

Aswin V is currently pursuing his Master of Technology degree in Artificial Intelligence, from Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India. He had finished his Bachelors from Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India, with a Bachelor of Technology degree in Electronics and Communication Engineering. He is passionate about Artificial Intelligence and Machine Learning applications. He had worked in a number of companies, both full-time and as an intern acquiring an assorted technical skill-set.

Vishnu Narayanan S is currently pursuing his Master of Technology degree in Artificial Intelligence, from Amrita Vishwa Vidyapeetham, Coimbatore, Tamil Nadu, India. He had finished his Bachelors from Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India with a degree of Bachelor of Technology in Computer Science and Engineering. He is passionate about Machine Learning and Deep Learning concepts, with a keen interest in computer applications.