IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 March 2024, accepted 14 May 2024, date of publication 17 May 2024, date of current version 28 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402350

RESEARCH ARTICLE

Investigating Gender and Age Variability in Diabetes Prediction: A Multi-Model Ensemble Learning Approach

RISHI JAIN[1,2], NITIN KUMAR TRIPATHI[1], MILLIE PANT[2], CHUTIPORN ANUTARIYA[1], AND CHAKLAM SILPASUWANCHAI[1]

[1]Asian Institute of Technology, Khlong Nueng 12120, Thailand
[2]Indian Institute of Technology Roorkee, Roorkee 247667, India

Corresponding author: Rishi Jain (st122603@ait.asia)

This work was supported in part by the Research Grant from Asian Institute of Technology, Thailand.

ABSTRACT The study investigates the intricate influence of gender and age variability in individuals diagnosed with diabetes, aiming to gain a comprehensive understanding of the diverse impact and implications of this prevalent metabolic disorder. A real-world dataset, obtained from a renowned diabetologist and meticulously maintained by Dr. Reddys' Lab, serves as the foundation for rigorous analysis. Leveraging the capabilities of ensemble learning, an advanced technique that combines multiple models, the predictive model's efficiency is substantially enhanced, resulting in precise and reliable predictions of individuals' diabetic status. Addressing the challenge of diabetes prediction, a novel ensemble learning model was proposed. The model combines the strengths of three distinct algorithms: Random Forest, Extra Trees, and Multilayer Perceptron (MLP). The model's output comprises a ternary label categorizing individuals as ''diabetic, non-diabetic, or pre-diabetic'', while the accompanying prediction score quantifies the likelihood of individuals belonging to each respective category. The findings of this research expand the existing body of knowledge on diabetes prediction, underscoring the untapped potential of ensemble learning methodologies in augmenting accuracy and predictive performance for diabetic patients.

INDEX TERMS Age groups, diabetes prediction, ensemble learning, gender variability, soft voting classifier.

I. INTRODUCTION

Diabetes, a chronic condition characterized by impaired glucose metabolism, poses a significant global health challenge. Diabetes prediction has been a subject of significant research interest, as highlighted in studies such as [1], [2], and [3]. These studies emphasize the overall importance of accurately predicting diabetes to enable early intervention and management strategies.

With its rising prevalence, understanding the various factors that influence the development, progression, and management of diabetes becomes increasingly crucial [4]. Among these factors, gender and age have emerged as important determinants contributing to the disease's diverse impact. Exploring the influence of gender and age variability on diabetic patients is essential for healthcare professionals, researchers, and policymakers to develop targeted interventions and strategies that address the unique needs and challenges faced by individuals with diabetes across different demographic groups.

Gender plays a notable role in diabetes, encompassing biological differences and social and cultural factors [5], [6]. Research suggests that the prevalence of diabetes differs between men and women [7], with variations in risk factors, disease progression, and complications. For instance, men may be more susceptible to certain risk factors such as central obesity, while women may face a higher risk of cardiovascular complications. Understanding these gender-specific differences can lead to tailored approaches in prevention, diagnosis, and treatment that take into account the specific vulnerabilities and challenges faced by each gender.

The associate editor coordinating the review of this manuscript and approving it for publication was Joanna Kołodziej.

Age, another critical factor, influences the onset, management, and outcomes of diabetes [8]. Diabetes manifests differently across age groups, with type 1 diabetes commonly diagnosed in children and young adults, while type 2 diabetes often occurs later in life. Age-related physiological changes, lifestyle factors, and comorbidities all contribute to the unique challenges faced by individuals with diabetes. Older adults [9], [10] for example, may struggle with self-care practices, medication adherence, and the management of multiple chronic conditions [11], necessitating specialized approaches to support their specific needs.

Exploring the influence of gender and age on diabetic patients goes beyond understanding the differences in disease prevalence or risk factors. It involves delving into the underlying biological mechanisms, genetic predispositions, hormonal variations, and behavioural patterns that contribute to the diverse impact of diabetes among different gender and age groups.

In investigating the relationship between diabetes and age and gender, authors in [8] and [12] have made noteworthy contributions. Their research indicates that age and gender play crucial roles in the development and prevalence of diabetes. Authors in [5] further support the association between age, gender, and diabetes. Their findings suggest that age and sex are one of the most influential factors related to both diabetes and pre-diabetes. Understanding how these factors influence diabetes risk can lead to more targeted interventions and personalized treatment approaches.

Moreover, authors in [13] have highlighted the significance of ensemble learning models over base models in the context of diabetes prediction. Ensemble learning combines the predictions of multiple individual models, resulting in improved accuracy and robustness. Their study demonstrates that ensemble models outperform standalone models, providing more reliable predictions and enhancing the effectiveness of diabetes prediction systems.

In this research, we address the complex issue of diabetes prediction, differentiating ourselves from previous studies by not only classifying individuals as diabetic or non-diabetic but also recognizing the crucial pre-diabetic state, a pivotal opportunity for early intervention. Moreover, we focus on specific demographic subgroups, considering age and gender, to tailor prevention and intervention strategies.

We employ ensemble learning, combining Random Forest, Extra Trees, and MLP models, a departure from the single-model approach prevalent in prior research. By doing so, we significantly enhance predictive accuracy, harnessing the strengths of these diverse models and thus advancing the precision of diabetes prediction. Additionally, we differentiate ourselves from most of the previous studies by employing a real-world dataset, providing a more accurate representation of the complexities present in the clinical setting. A unique aspect of this investigation lies in its consideration of different age groups (millennials, 30-40 age group) and genders (male, female), classifying individuals into diabetic, non-diabetic, or pre-diabetic categories with associated probability or likelihood scores. The potential applications of this research are diverse, ranging from assisting clinical decision-making and enabling personalized medicine to contributing to preventive healthcare strategies. Moreover, the nuanced analysis across demographics enhances the relevance and specificity of the predictive models. The research makes noteworthy contributions to predictive analytics methodology in healthcare, highlighting innovation through ensemble methods and nuanced demographic considerations, promising improvements in diabetes prediction and patient outcomes.

The study's scope encompasses proposing a novel ensemble model that integrates three distinct algorithms for enhanced predictive accuracy. The analysis is grounded in a real-world dataset, ensuring practical applicability. The model aims to categorize individuals into diabetic, non-diabetic, or pre-diabetic categories, offering a more refined approach to diabetes prediction. The study also delves into the importance of feature analysis, especially related to gender, and explores age-specific patterns to tailor predictions to different age groups.

The study explores the complex relationship between gender, age, and diabetes, utilizing a real-world dataset maintained by Dr. Reddys' Lab [14]. The proposed model can help doctors in providing a clear understanding of patient potential diabetes risk and can implement personalized care plans. The risk of delayed detection and intervention, leading to missed opportunities for early preventive measures can be mitigated. The AI-developed support system can also help reduce the time taken by doctors to diagnose patients.

This research seeks to advance predictive analytics in healthcare, specifically focusing on diabetes classification. A key aspect of this study involves a thorough comparison of various machine learning and deep learning models, leading to the careful selection of three optimal base models. Subsequently, these chosen models undergo a meticulous hyperparameter tuning process to improve their performance and generalizability. The comparison of various classifiers has shown that the soft voting classifier consistently demonstrated the highest performance, followed by the stacking classifier, then bagging, and finally, boosting classifiers across all gender and age scenarios. Then the chosen soft voting classifier ensemble model is applied to a real-world dataset, grounding the research in practical application. Below mentioned are the study highlights:

a) Exploration of Gender and Age Variability:

The study systematically investigates the intricate influence of gender and age on individuals diagnosed with diabetes. The research aims to provide a more comprehensive understanding of the diverse impact and implications of diabetes, recognizing the potential variations in predictive factors across different groups and gender.

b) Utilization of Real-World Dataset:

The research leverages a real-world dataset sourced from Dr. K D Modi, a renowned diabetologist. The maintenance of the dataset provides a robust foundation for the analysis, ensuring that the findings are grounded in authentic clinical

TABLE 1. Diabetes related research using machine learning.

	Description	Data	Methods	Results	Evaluation
Tasin et al. 2023	A machine learning-based automated diabetes prediction system that forecasts diabetes in female patients in Bangladesh	Private dataset of female patients in Bangladesh	Decision trees, SVM, Random forests, KNN, logistic regression, and ensemble methods. ADASYN.	GBoost classifier used the ADASYN technique, with an 81% accuracy rate, an F1 coefficient of 0.81, and an AUC of 0.84.	81% accuracy, F1 score 0.81, and an AUC of 0.84.
Su et al. 2023	Age adaption models for the risk prediction of diabetes mellitus	Pima Indian diabetes dataset	Neural networks, Support Vector Machines, Random Forest, Polynomial, Linear, and Logistic Regression, and XGBoost.	Linear regression has performed the best	77.7% accuracy, 80.1 % Precision, 77.8 Recall, and 78.9 F1 score
Dutta et al. 2022	An Ensemble of Machine Learning Models for the Early Prediction of Diabetes	Introduce a newly labelled diabetes dataset from Bangladesh.	A weighted ensemble of machine learning (ML) classifiers, including Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), XGBoost (XGB), and LightGBM (LGB).	Weighted_ensemble (DT + RF + XGB + LGB),	Accuracy of 0.735 and AUC of 0.832.
Abdalrada et al. 2022	Retrospective cohort research using machine learning models to predict the co-occurrence of diabetes and cardiovascular disease.	The dataset used is DiScRi (Diabetes complications screening research initiative).	A two-stage machine learning (ML) model to Forecast Diabetes Mellites co-occurrence.	Deep breathing heart rate changes, lying to standing blood pressure changes, HbA1c, HDL, Gender, Family History, and TC/HDL ratio were prominent factors	----
Mahesh et al. 2022	Improved chronic diabetes disease prediction and treatment using a blended ensemble model	Pima Indian diabetes dataset	Radial Basis Functions and Bayesian Networks were used.	The ensembled model performed better.	Accuracy of 97.11 %.
Deberneh & Kim 2021	Type 2 diabetes prediction using machine learning	Dataset was gathered using HER from 2013 to 2018	ANOVA tests, chi-squared tests, and recursive feature elimination method	The ensemble model performed better when compared to base models.	Accuracy of 82%

data, and enhancing the credibility and applicability of the research outcomes.

c) Ensemble Learning Approach:

The study employs ensemble learning, an advanced technique that combines the strengths of multiple models to enhance predictive efficiency. By integrating Random Forest, Extra Trees, and Multilayer Perceptron (MLP), the proposed ensemble model leverages diverse algorithmic capabilities, resulting in a more robust and accurate diabetes prediction system.

d) Outcome Classification and Likelihood Score:

The model's output, providing a ternary label categorizing individuals as ''diabetic,'' ''non-diabetic,'' or ''pre-diabetic,'' along with a prediction score, contributes to a more nuanced understanding of diabetes likelihood. This dual output adds depth to the predictive capabilities, enabling a finer classification of individuals and quantifying the certainty associated with each prediction.

In summary, this study investigates the influence of gender and age variability in individuals with diabetes, aiming to comprehensively comprehend the diverse impact and implications of the disease. By employing an ensemble learning technique, the model's efficiency is improved, resulting in accurate predictions of diabetic status, The output of the model is a ternary label indicating whether an individual is classified as diabetic, non-diabetic, or pre-diabetic. Additionally, the model provides a prediction score, quantifying the likelihood of an individual falling into each category.

II. LITERATURE REVIEW

The literature review in TABLE 1 and TABLE 2, provides a comprehensive overview of the diverse approaches employed by various authors in predicting diabetes within specific age groups or genders, encompassing both machine learning and statistical methodologies.

Authors in [15] proposed their work on predicting diabetes in female patients in Bangladesh employed a variety of machine-learning techniques. GBoost classifier with the

ADASYN approach for balancing the dataset performed the best, with an accuracy of 81% when compared to other models used for prediction. The authors in [16] performed a comparison of various ML models in diabetes prediction, and the results showed that Linear regression performed the best when the PIMA dataset was considered.

Ensemble learning helps in improving accuracy, enhancing robustness and stability and reducing overfitting. The authors in [17] and [18] proposed an *ensemble learning technique* in which various machine-learning models were used. The authors in [18] suggested a voting classifier ensemble combination of SVM+ANN models. When applied to the UCI dataset, this ensemble model achieved an impressive accuracy of 94.6%. Likewise, the authors in [19] highlighted the utilization of an ensemble model, comparing various combinations of models and base models. They determined that the combination of DT+RF+XGB+LGB, employing the weighted ensemble technique approach, yielded the most favourable performance (73.5% accuracy, 0.823 AUC). The studies concluded that the ensemble learning model performs better when compared to base models in classifying diabetic patients.

In order to perform reliable prediction, it is important to know what factors are to be considered, feature importance can play a vital role in deciding the variables that are to be considered for analysis purposes. Authors in [20] found that family history, gender, change in heart rate during deep breathing, change in blood pressure from lying to standing, HbA1c, HDL, and TC/HDL ratio were found to be prominent risk factors. Waist circumference and BMI values are significant factors, thus, to know their importance the authors in [21] performed an analysis and found that BMI had greater importance, among the BMI Asian/Hongkong criteria were prominent in predicting diabetes when compared to other nations. Apart from these hand grip strength [22] was also found to be one of the prominent factors in predicting diabetes in old age people.

In order to investigate the gender and age-specific differences in diabetes mortality, the authors in [12] evaluated correlations between age- and gender-specific diabetes mortality using negative binomial regression. It was seen that women had higher mortality than men and people above 75 years are vulnerable. Investigating the relationship between age groups with diabetes authors in [9] found that in China, middle-aged and older adults have a significant risk of developing diabetes, whereas [23], [24] also obtained similar results in which the risk of diabetes was found to be high in adults.

The reviewed literature provides valuable insights into the prediction, risk factors, and demographic associations related to diabetes. Several machine-learning techniques have been explored, with GBoost classifier and ensemble learning demonstrating promising results in accurately classifying diabetes. Key risk factors such as family history, gender, physiological measurements, and BMI have been identified, shedding light on the importance of these variables in diabetes prediction. Furthermore, the studies highlight age and gender-specific variations in diabetes mortality, indicating higher vulnerability among older individuals and women.

TABLE 2. Related works pertaining to statistical methodologies.

	Description	Data	Methods	Results
Ying Tian et al. 2023	The association between diabetes and physical exercise in middle-aged and older adults.	Charles_data set (2018)	The Z-test, linear hierarchical regression, and logistic regression.	Middle-aged and older chinese citizens were find to be at major risk.
Chen et al. 2022	Diabetes and the effect of older age adiposity	Interviewed 2,608 participants	The Cox regression model	Compared to other BMI criterion, the BMI-Asian/Hong Kong criteria were found to more prone.
Zheng et al. 2022	Age- and Gender-Specific variations in Shandong, China	China National Death Surveillance System data	Negative binomial regression	Men were less likely to die than women, and those over 75 were the most vulnerable.
Cai et al. 2021	Prediction model to forecast the Type 2 diabetes incidence.	Data was obtained from 12,940 non-obese individuals over 5 years without diabetes	A multivariate Cox regression analysis	A1c glycosylated haemoglobin, age, fatty liver, γ-glutamyl transpeptidase, triglycerides, and fasting plasma glucose were risk factors.
Zhang et al. 2021	Diabetes's seasonal, gender, and geographic variations.	Sixth Chinese population census (2010)	X^2 test and standardized statistics	Higher blood pressure, higher serum total cholesterol levels, higher BMI, and older ages were prominent risk factor found.
Kunutsor et al. 2020	Strengthening of the handgrip enhances type-2 diabetes prediction	The dataset contains 776 people, between the ages of 60 and 72, who have no prior history of T2D.	Integrated discrimination index (IDI) and Net reclassification improvement (NRI)	Handgrip strength is considered to be one of the prominent factors in predicting diabetes in women.

The risk of diabetes has been consistently observed to be elevated in middle-aged and elderly populations.

The existing literature has highlighted key risk factors associated with diabetes, encompassing elements such as family history, gender, physiological measurements, BMI and overall diabetes prediction. While age is acknowledged as a factor, the literature lacks a comprehensive examination of age as a variable, which plays a pivotal role in diabetes risk assessment. Furthermore, the absence of detailed consideration for specific markers, such as Hemoglobin A1c (hBa1c), fasting glucose (GTT0), triglycerides (TG), uric acid levels, systolic and diastolic blood pressure (SYS BP, DIA BP), total cholesterol (T. CHOL), low-density lipoprotein (LDL), high-density lipoprotein (HDL), and non-high-density lipoprotein (NON-HDL), indicates an opportunity for an exploration of biochemical indicators. Addressing these gaps in the literature is essential for a clear understanding of the multifaceted factors contributing to diabetes and for refining predictive models for enhanced accuracy.

The proposed AI-developed model can help doctors in providing a clear understanding of patient potential diabetes risk and can implement personalized care plans. The risk of delayed detection and intervention, leading to missed opportunities for early preventive measures can be mitigated. The AI-developed support system can also help reduce the time taken by doctors to diagnose patients.

III. DATA DESCRIPTION

A dataset related to diabetes was collected from a diabetologist and maintained by Dr. Reddys' lab. As shown in Table 3 the dataset contains 13 independent variables and one dependent variable, namely "outcome". The independent variables are gender, age, hba1c, BMI, gtt0(fasting glucose), LDL, HDL, nonhdl, tg, uric acid, sys bp, and dia bp.

Gender is a categorical variable that can have two values: male or female. Age is a continuous variable that represents the patient's age in years. Hba1c is a continuous variable that represents the patient's blood sugar level. BMI is a continuous variable that represents the patient's body mass index. Gtt0 (fasting glucose) is a continuous variable that represents the patient's blood glucose level after fasting for 8 hours.

LDL, HDL, nonhdl, tg, and uric acid are all continuous variables that represent different types of lipids in the patient's blood. Sys bp and dia bp are continuous variables that represent the patient's systolic and diastolic blood pressure, respectively.

The dependent variable, "outcome", is a categorical variable that has three values: non-diabetic, pre-diabetic, and diabetic. The outcome is determined based on the patient's results from a glucose tolerance test (GTT), which is considered to be the gold standard for diagnosing diabetes.

Glucose Tolerance Test: A GTT involves fasting for 8 hours and then drinking a sugary drink. The patient's blood glucose level is then measured at regular intervals over the next 2 hours. If the patient's blood glucose level is high after drinking the sugary drink, they may have diabetes.

TABLE 3. Features used and their description.

Feature	Description
Gender	Male or Female
Age	Age of the person
hBa1c	Hemoglobin A1c
BMI	Body mass index
GTT0	Fasting glucose
TG	Triglycerides
Uric Acid	Level of uric acid in the bloodstream
SYS BP	Systolic blood pressure
DIA BP	Diastolic blood pressure
T. CHOL	Total cholesterol
LDL	Low-density lipoprotein
HDL	High-density lipoprotein
NON-HDL	Non-High-density lipoprotein
Outcome	Non-diabetic (0), Pre-diabetic (1), Diabetic (2)

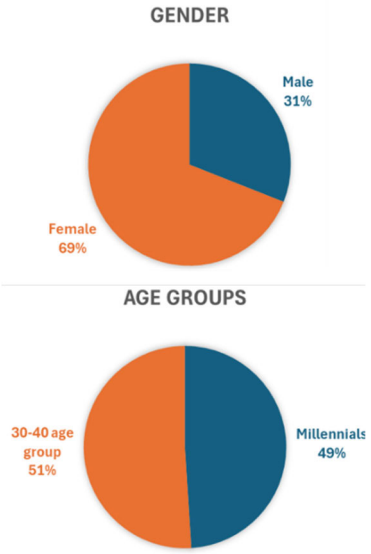


FIGURE 1. Gender and age-group distribution in the dataset.

The dataset comprises information on individuals belonging to two distinct age groups: 30-40 years and millennials. Notably, there are 311 tuples corresponding to the 30-40 age group and 297 tuples pertaining to the millennial cohort as depicted in FIGURE 1. The specialist, leveraging his clinical expertise and the gold standard glucose tolerance test, has proficiently assigned outcome variables to each tuple, classifying them into three categories: diabetic, non-diabetic, and pre-diabetic.

Gender and age-group distribution in the dataset.

We've categorized our dataset according to GTT values: $\geq$200 mg/dL indicates diabetes, 140-199 mg/dL signals pre-diabetes, and $<$140 mg/dL denotes normal glucose levels, as defined by the CDC [25].

Furthermore, the dataset demonstrates a notable gender distribution, with a total of 461 instances attributed to females and 211 instances associated with males. This gender and age-disaggregated data presents an avenue for conducting gender-based and age-group-based analyses in the context of diabetes, allowing for the examination of potential variations

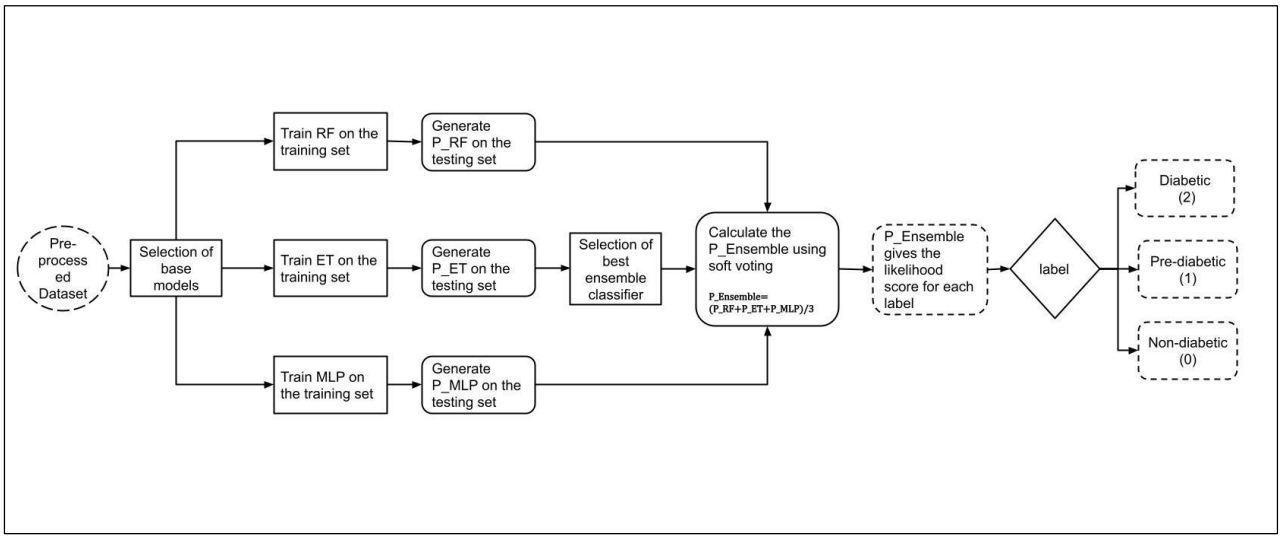


FIGURE 2. Flowchart of proposed methodology.

and trends in disease prevalence and outcomes between different genders and age groups.

IV. METHODOLOGY

The flowchart in Figure 2 demonstrates steps taken for building an ensemble model. It involves data collection, preprocessing, selection and application of ensemble techniques and obtaining diabetes classification and probability scores.

a) Data Collection:

Acquiring a real-world dataset from Care Hospital (Dr. K.D Modi), Hyderabad maintained by Dr. Reddyss' Lab The dataset encompasses information related to various health indicators, including gender, age, HbA1c levels, BMI, fasting glucose (gtt0), LDL cholesterol, HDL cholesterol, non-HDL cholesterol, triglyceride levels (tg), uric acid levels, systolic blood pressure (sys bp), and diastolic blood pressure (dia bp). All the permissions have been taken to use the data.

b) Data Preprocessing:

Cleaning and preprocessing the dataset to handle missing values, and outliers, and standardize features to ensure data quality.

c) Ensemble Model Training:

- Implementing the ensemble learning approach using three distinct algorithms: Random Forest, Extra Trees, and Multilayer Perceptron (MLP).
- Training each base model on the pre-processed dataset to harness their individual predictive strengths.
- Comparing various ensemble learning techniques and choosing the best classifier

d) Soft Voting Classifier:

Employing a soft voting classifier technique to aggregate predictions from the three base models, producing a final ensemble prediction.

e) Likelihood Score Calculation:

Generating a prediction score for each individual, quantifying the likelihood of belonging to each respective diabetic category.

f) Diabetes Classification:

Categorizing individuals into ''diabetic,'' ''non-diabetic,'' or ''pre-diabetic'' based on the ensemble model's output.

A. DATA PREPROCESSING

1) DATA LABELLING

In the context of diabetes data, data labelling assumes a pivotal role, aided by the expert diagnoses provided by Dr. K.D Modi, a distinguished diabetologist in Hyderabad. Dr. Modi's assessments serve as an authoritative source for categorizing individuals into specific health states, where ''diabetes'' is encoded as ''2,'' ''pre-diabetic'' as ''1,'' and ''non-diabetic'' as ''0. This data labelling process ensures that each data point is accurately assigned a class label, thereby enabling the training and evaluation of machine learning models tailored for diabetes prediction. The labels were thus assigned.

2) DATA IMPUTATION

Dr. Modi's medical insights and advice were instrumental in filling in the gaps left by missing data, ensuring the completeness and accuracy of the dataset. The height and weight of females were replaced with 152 cm and 50 kgs, whereas for males it was 162 and 65 kg.

3) DATA AUGMENTATION

To mitigate the dataset's inherent imbalance, where women were overrepresented compared to men, the Synthetic Minority Over-sampling Technique (SMOTE) was effectively employed. SMOTE is a method used in machine learning to create synthetic instances of the minority class (in this case, the ''men'' category) by interpolating between existing data points. This technique helps balance the class distribution, ensuring that the model is not biased toward the majority class. By applying SMOTE, the dataset was rebalanced, allowing for a more equitable representation of both genders.

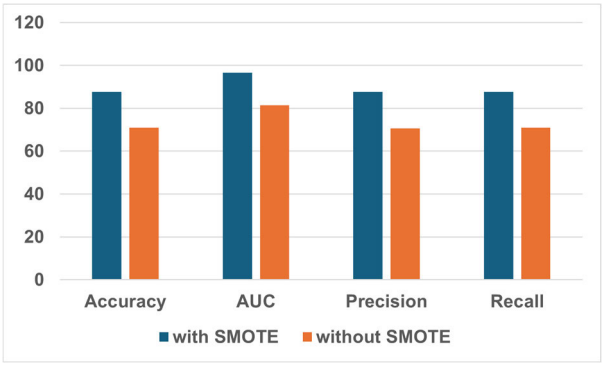


FIGURE 3. Importance of SMOTE technique and impact on results.

The Figure 3 shows the comparative analysis of the results obtained by the entire dataset with and without using the SMOTE technique. The SMOTE dataset in overall performs better when compared to the dataset in which SMOTE is not applied.

B. MODEL USED

1) RANDOM FOREST

Random Forest is a powerful algorithm used for classification tasks. It integrates the predictions of several decision trees to determine the final classification [26], [27]. Every decision tree splits the data at each node using a random subset of features after being trained on a random subset of the training set. The final projected class for a given input sample is assigned to the class that receives the most votes from each individual tree using majority voting technique.

2) EXTRA TREES

Every decision tree in Extra Trees is trained using a random subset of the training set, and feature selection is done using random splits based on random threshold values [28]. Extra Trees chooses splits at random without taking into account the best thresholds, in contrast to Random Forest, which chooses the best split from a subset of features. This increases the model's unpredictability and reduces its propensity for overfitting.

3) MULTI-LAYER PERCEPTRON

Multilayer Perceptron (MLP) is a popular artificial neural network architecture used for classification tasks [29], [30]. It is made up of several layers of networked nodes, often referred to as neurons, arranged into an input layer, an output layer, and one or more hidden layers.

C. ENSEMBLE LEARNING

Ensemble learning produces predictions that are more reliable and accurate than any one model could be by combining several base models [31], [32], [33]. To enhance the models' performance on intricate datasets, ensemble approaches are frequently employed.

TABLE 4. Base models and their respective parameters.

Algorithm	Hyper Parameters	Value
RF	*Randomly subsetset selection*	*Allow*
	Quality of split evaluation	*Gini* method
	no. of decision tree	100
	min. *no. of samples required to split*	2
	max. features selected	*auto*
MLP	*Activation function*	*Relu*
	Regularizer (α)	0.0001
	Optimizer	*Adam*
	no. of neurons	100
	no. of hidden layer	1
	max. *training epochs*	500
	momentum	0.9
	initial learning rate	0.001
ET	*Complete subsetset selection*	*Allow*
	Quality of split evaluation	*Gini method*
	no. of decision tree	100
	min. *no. of samples required to split*	2
	max. features selected	*auto*

The soft voting classifier is an ensemble learning method that generates a final prediction by aggregating the predicted probabilities of several base models [34], [35]. Every base model produces a probability distribution across the classes during soft voting. Then, a single probability distribution is created by combining these probability distributions. The final prediction is the class with the highest probability in the combined distribution.

The proposed ensemble model methodology takes advantage of the diverse strengths of Random Forest, Extra Trees, and MLP algorithms. Random Forest is a powerful algorithm that integrates the predictions of multiple decision trees, while Extra Trees further enhances unpredictability and reduces overfitting. MLP, on the other hand, is an artificial neural network architecture commonly used for classification tasks.

The ensemble model utilizes a soft voting classifier approach in which each base model produces a probability distribution across the classes, and these probability distributions are combined to create a single probability distribution. The final prediction is made based on the class with the highest probability in the combined distribution.

1) MATHEMATICAL MODELLING OF ENSEMBLE LEARNING TECHNIQUE

The ensemble learning technique employed in this study combines the strengths of three distinct algorithms: Random Forest, Extra Trees, and Multilayer Perceptron (MLP). Let's denote the predictions of these models as P_{RF}, P_{ET}, and P_{MLP} respectively. The mathematical representation of the ensemble prediction (P_{Ensemble}) can be expressed as follows:

$$P_{Ensemble} = \frac{(P_{RF} + P_{ET} + P_{MLP})}{3} \quad (1)$$

This ensemble model output provides a ternary label classifying individuals as "diabetic," "non-diabetic," or "pre-diabetic." Additionally, an associated prediction score is

Algorithm: Ensemble Diabetes Prediction Model

Input: Pre-processed Dataset (D)
Output: Ensemble Prediction (P_Ensemble)
1. Split D into training and testing sets.
2. Train Random Forest model (RF) on the training set.
3. Train Extra Trees model (ET) on the training set.
4. Train Multilayer Perceptron model (MLP) on the training set.
5. Generate predictions P_RF, P_ET, and P_MLP on the testing set.
6. Choosing the best ensemble technique
7. Calculate the ensemble prediction $P_{Ensemble}$ using soft voting:
$P_{Ensemble} = (P_{RF} + P_{ET} + P_{MLP})/3$
8. Calculate the likelihood score for each individual.
9. Classify individuals based on $P_{Ensemble}$ into "diabetic (2)," "non-diabetic (0)," or "pre-diabetic (1)."
10. Output the Ensemble Prediction and likelihood scores.

generated for each individual, quantifying the likelihood of belonging to each respective category.

A Pre-processed dataset D with labels L_0, L_1 and L_2 is divided into Training data X, validation data U and testing data V. The training data is subjected to N classifiers C_1, C_2……., C_{n-1}, C_n. Each of the classifiers will make a prediction P_1, P_2…., P_{n-1}, and P_n respectively. In the case of Hard voting also called Majority voting, if the majority of the classifier prediction is L_p then it can be confirmed that validation data belongs to the class L_p.

Whereas in the case of soft voting, each of the classifiers C will provide a probability for each of the labels. For example, C_1 classifier will generate probabilities $P_1(L_0)$, $P_1(L_1)$ and $P_1(L_2)$. Once the probabilities are obtained from all the classifiers for L_0, L_1 and L_2, then an average is taken for each of the label probabilities.

$$P(L0) = \frac{P_1(L_0) + P_2(L_0) + P_3(L_0) \ldots + P_{n-1}(L_0) + P_n(L_0)}{N} \tag{2}$$

$$P(L1) = \frac{P_1(L_1) + P_2(L_1) + P_3(L_1) \ldots + P_{n-1}(L_1) + P_n(L_1)}{N} \tag{3}$$

$$P(L2) = \frac{P_1(L_2) + P_2(L_2) + P_3(L_2) \ldots + P_{n-1}(L_2) + P_n(L_2)}{N} \tag{4}$$

Once the cumulative probability is obtained then if the probability of class '0' is greater than the cumulative probability of class '1' and class '2' then the dataset belongs to class '0'.

Soft voting is a more sophisticated approach than hard voting, and it can often lead to improved accuracy. This is because soft voting takes into account the uncertainty of the base models' predictions.

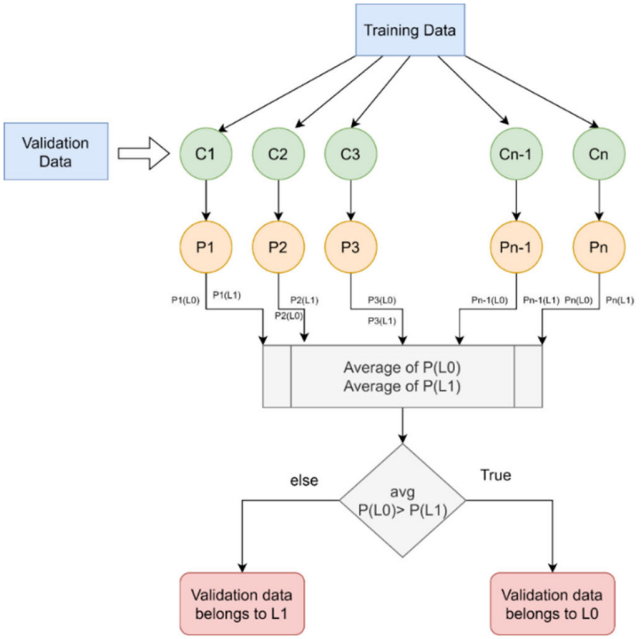


FIGURE 4. Soft voting classifier (ensemble technique).

The FIGURE 4 above shows the working of a soft voting classifier with two labels L1 and L2. By combining the probability distributions of the base models, soft voting is able to make a more informed decision about the final prediction.

In a nutshell, the proposed methodology depicted in the Figure 5 comprises of comparison of various machine learning models and deep learning models, followed by a selection of three best base model, then hypertuning the base model and then the hyper-tuned base model are subjected to soft voting classifier to create an ensembled model which is subjected to the real world dataset. The analysis is conducted considering various age groups (millennials, 30-40 age group) and gender (male, female) based scenarios which thus classify them as diabetic, non-diabetic or pre-diabetic, with a probability or likelihood score. Exploring these specific demographic segments, the research aims to understand how diabetes prediction varies across age and gender populations.

V. EXPERIMENTAL SETUP

The experimental setup section of the study contains information on the hardware and software components used, stratified k-fold cross validation and evaluation metrics. The hardware and software sub-section includes information on the system configuration, programming language and associated libraries used for building and running the proposed model.

A. HARDWARE AND SOFTWARE COMPONENTS

Hardware and software components help to define the performance, functionality, and adaptability of a computing system. Selecting the right combination of hardware and software is crucial for meeting specific computing needs efficiently

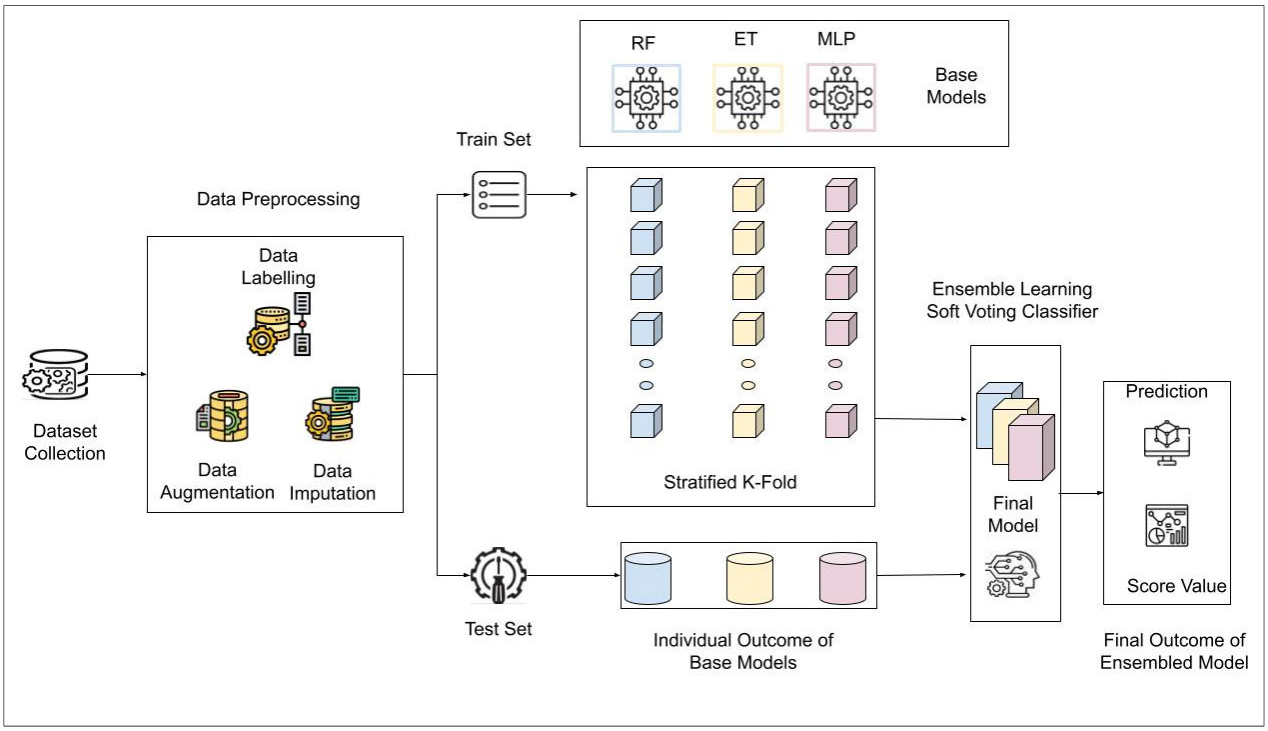


FIGURE 5. Proposed ensembled model.

TABLE 5. Hardware and software components used.

Processor	11th Gen Intel (R) Core (TM) i5-1155G7 @ 2.50 GHz
RAM	12.0 GB
Operating System	64-bit operating system
Programming Language	Python 3.9
Library	pycaret, plotly, seaborn

and effectively. The TABLE 5 describes the specifications of hardware and software components used for predicting diabetes.

B. PERFORMANCE EVALUATION

Stratified K-Fold: It is a type of cross-validation fold generator that ensures that the distribution of the target variable is similar in each fold [36]. It works by first creating a stratified split of the data, which means that the target variable is evenly distributed across the folds. This is done by first creating a list of all the unique values of the target variable. Then, for each fold, a random sample of data points is selected from each unique value of the target variable. This ensures that each fold has a similar distribution of the target variable as the overall population. 10-fold cross-validation was used to avoid overfitting. TABLE 6 shows the five evaluation metric's used to evaluate the performance of the model. Using the Stratified K-Fold cross-validation ensures that the distribution of the target variable is similar in each fold. Specifically, a 10-fold cross-validation method is used where the dataset is divided into 10 equal parts, and the model is trained and evaluated 10 times which would help in avoiding overfitting.

Accuracy: Percentage of instances that are correctly classified.

AUC: Estimating the area under the Receiver Operating Characteristic (AUC), a ternary classification model's overall performance is quantified.

Recall: The ratio of a model's real positive predictions to its total positive predictions

Precision: The ratio of true positives to all actual positives. It is often referred to as sensitivity or true positive rate.

F1 score: The harmonic mean of precision and recall

Five evaluation metrics were used namely accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics collectively provide a comprehensive analysis of the model's predictive efficacy.

VI. FEATURE IMPORTANCE

The mean decrease in impurity (MDI) measures how much a feature helps in making decisions in a tree-based model. It looks at how much the disorder or randomness in the data is reduced when the model uses that feature to split the data.

$$MDI = \sum \frac{(Impurity\ BS - Impurity\ AS)}{Number\ of\ Splits} \tag{5}$$

Before Split(BS): Check how messy the data is before using the feature to split.

After Split(AS): See how much cleaner the data becomes after the split.

Calculate the Reduction: Measure how much the disorder decreased.

TABLE 6. Evaluation metrics and their description.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
AUC	$\sum_{i=1}^{n} \frac{(FPR[i]-FPR[i-1])\cdot(TPR[i]+TPR[i-1])}{2}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1	$2\times\frac{Precision\times Recall}{Precision+Recall}$

Average Across Splits: Do this for all the times the model uses the feature to split, and find the average reduction.

The higher the mean decrease in impurity value for a feature, the more important it is considered to be in making predictions.

A. GENDER

In order to select the most important features, the ''Mean decrease in impurity method'' or ''Gini Index'' was used. This method calculates the average decrease in impurity of a decision tree when a particular feature is split [37]. Features that cause the greatest decrease in impurity are considered to be the most important. Based on the feature importance graph FIGURE 6, it is evident that for females, the most significant features are fasting glucose, HbA1c, Non-high-density lipoprotein (NONHDL), uric acid, and low-density lipoprotein (LDL), followed by systolic blood pressure.

Conversely, for males, the feature importance graph indicates that fasting glucose, triglycerides, HbA1c, total cholesterol, low-density lipoprotein (LDL), Non-high-density lipoprotein (NONHDL), and high-density lipoprotein (HDL) are the most influential features.

B. AGE GROUP

Based on the analysis of the feature importance graph FIGURE 7, it is noteworthy that fasting glucose, gender, HbA1c, triglycerides, body mass index, and total cholesterol are identified as the primary influential features for the millennial population

Conversely, for individuals in the age group of 30 to 40 years, the feature importance graph highlights fasting glucose, HbA1c, low-density lipoprotein (LDL), gender, total cholesterol, and diastolic blood pressure as the most crucial features.

VII. APPLICATION OF PROPOSED MODEL ON UCI REPOSITORY DATASET

There is a risk of delayed detection and intervention, leading to missed opportunities for early preventive measures.

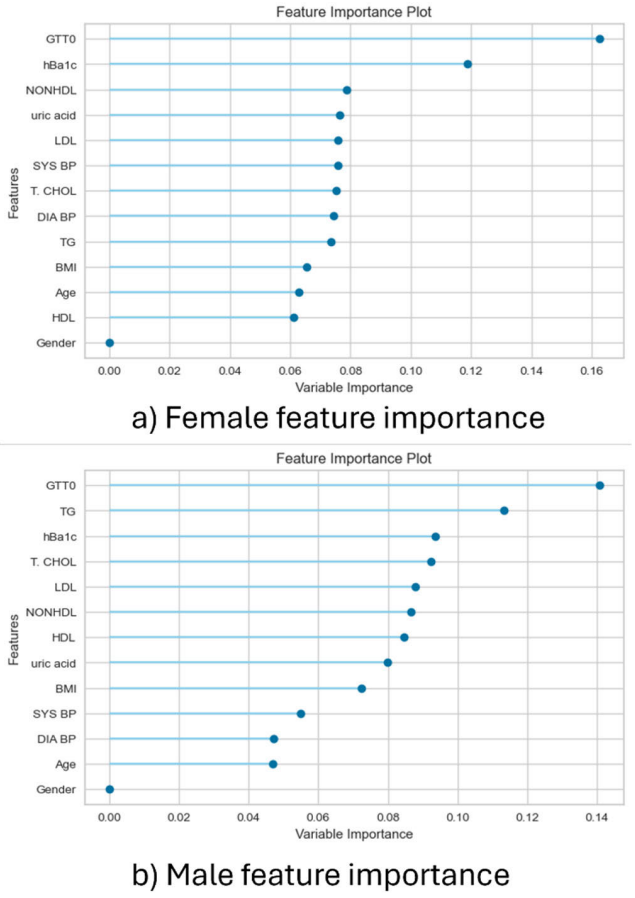
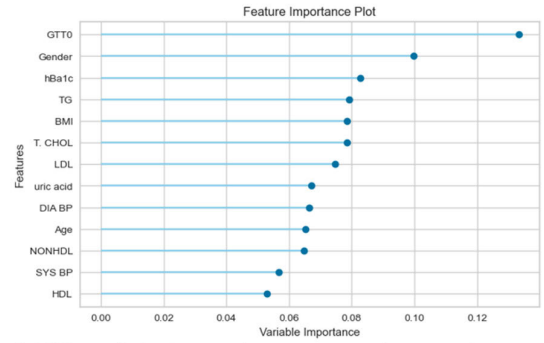


FIGURE 6. Graph representing feature importance values for a) female and b) male.

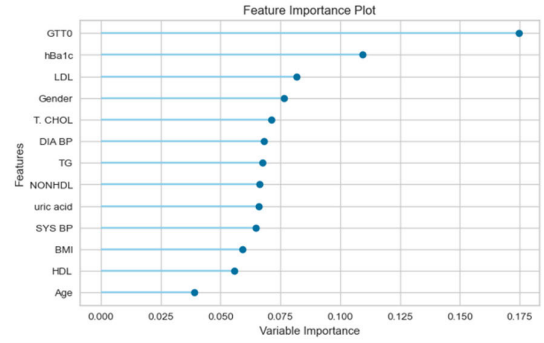
Patients may remain undiagnosed or improperly categorized, hindering the development of targeted treatment plans tailored to their specific risk profiles. The ability to categorize individuals as diabetic, non-diabetic, or pre-diabetic, accompanied by a likelihood score, offers substantial benefits to both healthcare providers and patients.

We have already taken a real-world dataset from Dr. K. D Modi, Care Hospital, Hyderabad, India maintained by Dr. Reddy's Lab. To showcase the generalizability of our proposed model we applied our ensemble proposed model to a widely known secondary dataset [1], [38], [39], [40]: Sylhet Diabetes Hospital in Sylhet, Bangladesh UCI data. We downloaded it from Kaggle [41]. The results in Figure 8 have shown that our model has performed pretty well on it showing an accuracy of 99.4% and 1.0 AUC. Thus, from this robustness of our model can be seen.

The proposed model and results were discussed with the doctor, and he mentioned that the proposed model can classify the overall risk status as per age and gender-wise distribution. It can be useful for doctors to target these patients for future cardiovascular and metabolic risks. This may include lifestyle modifications, early pharmacological interventions, or referrals to specialized diabetes care programs.



a) Millennials (18-30) age group feature importance



b) 30-40 age group feature importance

FIGURE 7. Graph representing feature importance values for a) Millennials and b) Adults (30-40 years).

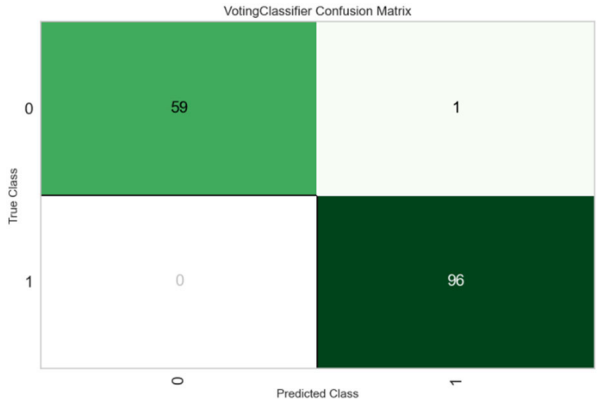


FIGURE 8. UCI data confusion matrix.

VIII. RESULTS AND DISCUSSION

A. COMPARISON OF VARIOUS ENSEMBLE CLASSIFIERS

A comparative analysis is conducted by applying bagging, boosting, stacking, and soft-voting classifiers to the three base models chosen. The findings from the bar chart and heat map Figure 9 indicate that the soft voting classifier consistently demonstrated the highest performance, followed by the stacking classifier, then bagging, and finally, boosting classifiers across all gender and age scenarios. Thus, based on the analysis obtained, we concluded that the soft voting classifiers ensemble technique performed better.

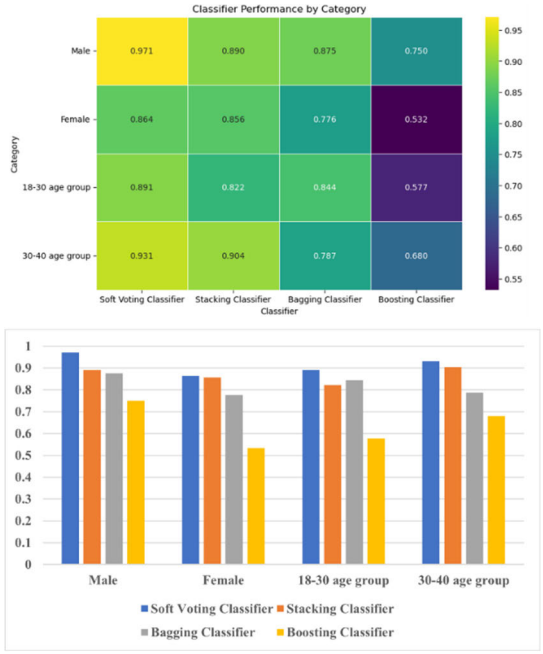


FIGURE 9. Comparison of various ensemble technique classifiers.

TABLE 7. Evaluation metrices and values obtained for "Female".

Evaluation Metrics	Value
Accuracy	0.864
AUC	0.970
Recall	0.864
Precision	0.854
F1	0.860

FIGURE 10. Confusion matrix obtained for "female".

B. GENDER (FEMALE)

The TABLE 7 depicts the evaluation metrics value obtained for female category after applying the proposed ensembled model.

a) Confusion Matrix

It is evident from FIGURE 10 that the proposed ensemble model performs best in the case of predicting diabetes (2), followed by pre-diabetic (1) and non-diabetic (0) respectively.

b) AUC and Precision and Recall

It can be seen from Figure 11, the classifier is slightly less accurate in predicting class 0, with an AUC value of 0.94. This means that the classifier correctly predicts the class of 94% of all samples from class 0. The classifier is also

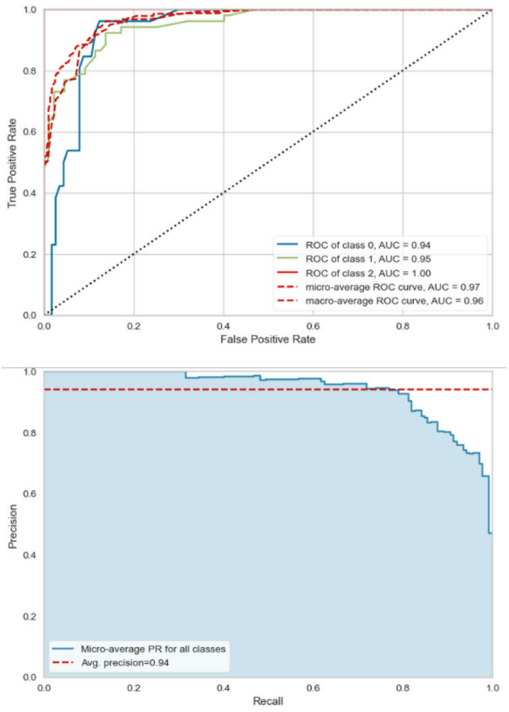


FIGURE 11. AUC and Precision- Recall curves obtained for "female".

very accurate in predicting class 1, with an AUC value of 0.95. This means that the classifier correctly predicts the class of 95% of all samples from class 1. The classifier is most accurate in predicting class 2, with an AUC value of 1.00. This means that the classifier correctly predicts the class of 100% of all samples from class 2.

The PR graph shows that the classifier is able to achieve high precision and recall values. This means that the classifier is able to correctly predict the class of most samples, while also avoiding false positives. The classifier is able to avoid false positives 94% of the time from the PR graph.

c) Class Prediction Error

A "class prediction error" refers to an occurrence when a model inaccurately assigns a data point to the wrong category or fails to correctly identify the class it belongs to during classification or prediction. In other words, it is an error in which the model's prediction does not align with the actual class or category of the data point.

From the graphs Figure 12, it can clearly be seen that the model classifies diabetic (2) people most accurately as it has the least prediction error, followed by non-diabetic (0) and lastly pre-diabetic (1). The blue bar is the tallest, which means that the model is most likely to misclassify samples from class 1.

d) Type-1 and Type-2 Error

$$Type\ 1\ error = \frac{FP}{(FP + TN)} \tag{6}$$

$$Type\ 2\ error = \frac{FN}{(FN + TP)} \tag{7}$$

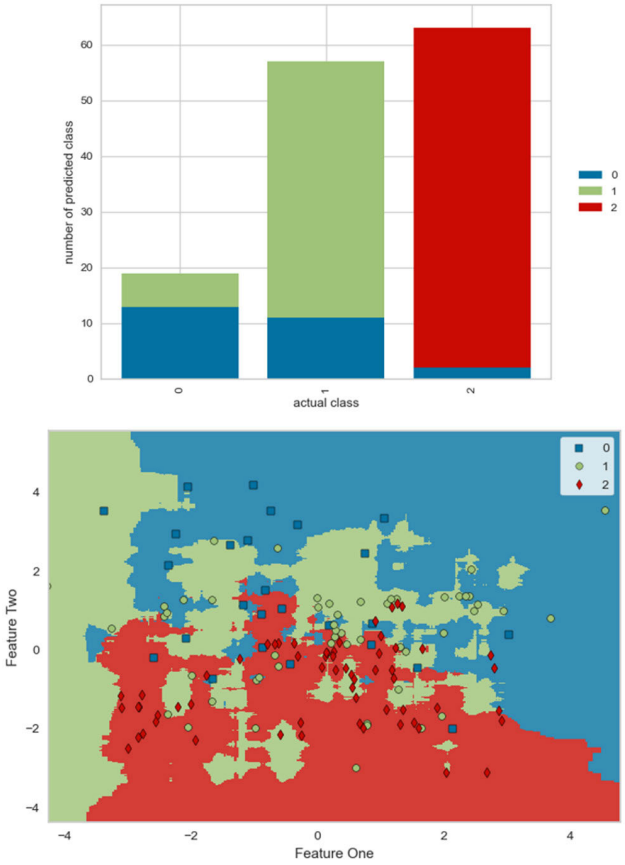


FIGURE 12. Class prediction error graph obtained for "female" where 0:non-diabetic, 1: pre-diabetic and 2: diabetic.

TABLE 8. Type-1 and Type-2 error values obtained for "Female".

Female	Type-1 Error	Type-2 Error
Overall	0.068	0.205
Diabetic	0.025	0
Pre-diabetic	0.126	0.115
Non-diabetic	0.053	0.5

The TABLE 8 and Figure 13, provides information on the Type-1 and Type-2 error rates for females across different diabetic categories. In the overall female population, the Type-1 error rate is 0.068, indicating a moderate occurrence of false positive diagnoses. The Type-2 error rate is higher at 0.205, suggesting a substantial likelihood of false negative diagnoses. Among females identified as diabetic, the Type-1 error rate is 0.025, reflecting a relatively low probability of false positive diagnoses. Interestingly, there were no false negative diagnoses observed for diabetic females. For females classified as pre-diabetic, the Type-1 error rate is significantly higher at 0.126, indicating a substantial chance of false positive diagnoses, while the Type-2 error rate stands at 0.115, suggesting a moderate likelihood of false negative diagnoses. Notably, among females identified as non-diabetic, the

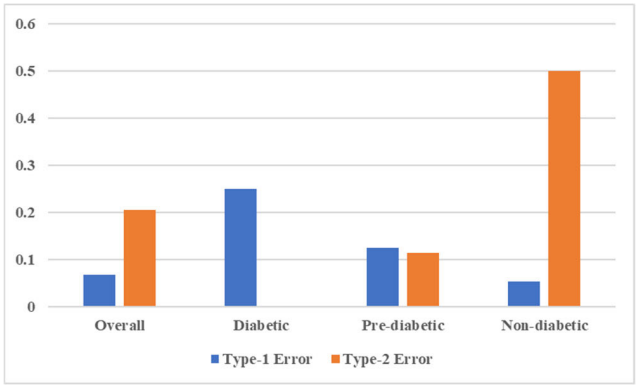


FIGURE 13. Type-1 and Type-2 error graph obtained for "female."

TABLE 9. Evaluation metrices and values obtained for "male".

Evaluation Metrics	Value
Accuracy	0.971
AUC	0.991
Recall	0.974
Precision	0.973
F1	0.970

FIGURE 14. Confusion matrix obtained for "male".

Type-1 error rate is 0.053, signifying a moderate occurrence of false positive diagnoses, while the Type-2 error rate is notably higher at 0.5, indicating a high probability of false negative diagnoses.

C. GENDER (MALE)

The TABLE 9 depicts the evaluation metrics value obtained for male category after applying the proposed ensembled model.

a) Confusion Matrix

It is evident from Figure 14, that the proposed ensemble model performs best in the case of predicting diabetes (2), followed by pre-diabetic (1) and non-diabetic (0) respectively.

b) AUC and Precision and Recall

It can be seen from Figure 15 the classifier is most accurate in predicting class 0, with an AUC value of 1.00. This means that the classifier correctly predicts the class of 100% of all samples from class 0. The classifier is also very accurate in predicting class 1, with an AUC value of 0.99. This means that the classifier correctly predicts the class of 99% of all samples from class 1. The classifier is slightly less accurate in

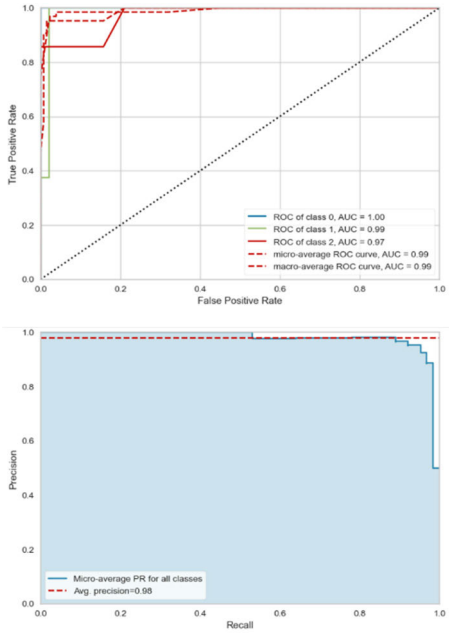


FIGURE 15. AUC and Precision-Recall curves obtained for "male".

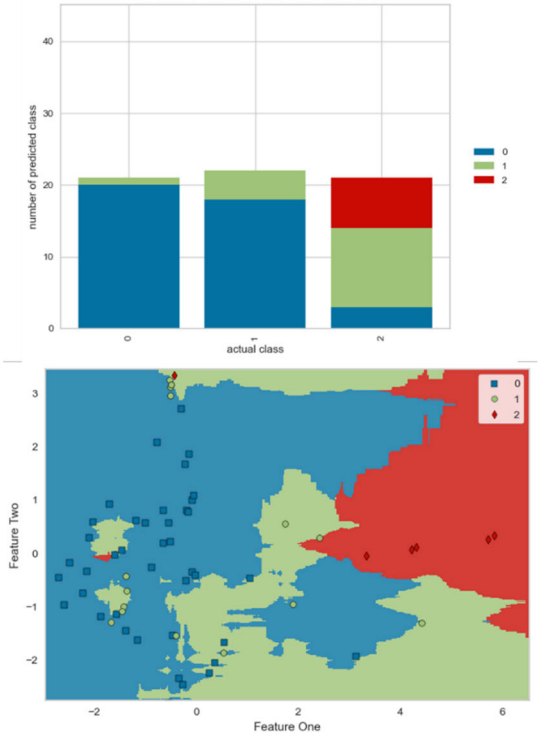


FIGURE 16. Class prediction error graph obtained for "male" where 0:non-diabetic, 1: pre-diabetic and 2: diabetic.

predicting class 2, with an AUC value of 0.97. This means that the classifier correctly predicts the class of 97% of all samples from class 2. The classifier is able to avoid false positives 94% of the time from PR graph.

c) Class Prediction Error

From the above graph Figure 16, it can clearly be seen that the model classifies non-diabetic (0) people most accurately

TABLE 10. Type-1 and Type-2 error values obtained for "Male".

Male	Type-1 error	Type-2 error
Overall	0.021	0.068
Diabetic	0	0.142
Pre-diabetic	0.02	0.062
Non-diabetic	0.043	0

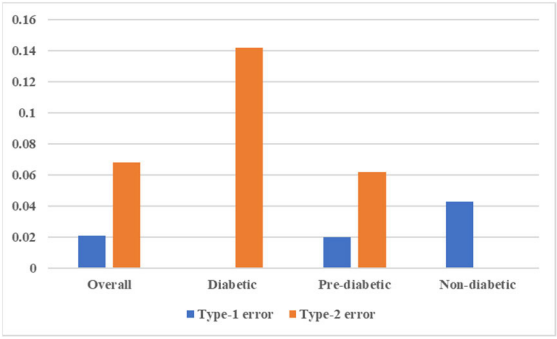


FIGURE 17. Type-1 and Type-2 error graph obtained for "male."

TABLE 11. Evaluation metrices values obtained for "Millennials."

Evaluation Metrics	Value
Accuracy	0.891
AUC	0.971
Recall	0.891
Precision	0.891
F1	0.891

as it has the least prediction error, followed by pre-diabetic (1) and lastly diabetic (2).

d) Type-1 and Type-2 Error

The provided TABLE 10 and Figure 17 presents the Type-1 and Type-2 error rates for males across different diabetic categories. In the overall male population, the Type-1 error rate is recorded as 0.021, indicating a relatively low occurrence of false positive diagnoses. However, the Type-2 error rate is higher at 0.068, suggesting a moderate likelihood of false negative diagnoses. For males identified as diabetic, both the Type-1 and Type-2 error rates are observed to be 0, signifying an absence of false positive and false negative diagnoses. Among males classified as pre-diabetic, the Type-1 error rate is 0.02, indicating a minor probability of false positive diagnoses, while the Type-2 error rate stands at 0.062, reflecting a moderate chance of false negative diagnoses. Notably, for males identified as non-diabetic, the Type-1 error rate is 0.043, indicating a moderate likelihood of false positive diagnoses, while the Type-2 error rate is recorded as 0, denoting an absence of false negative diagnoses.

D. AGE INTERVAL ANALYSIS (18- 30 (MILLENNIALS))

The TABLE 11 depicts the evaluation metrics value obtained for millennials age group after applying the proposed ensembled model.

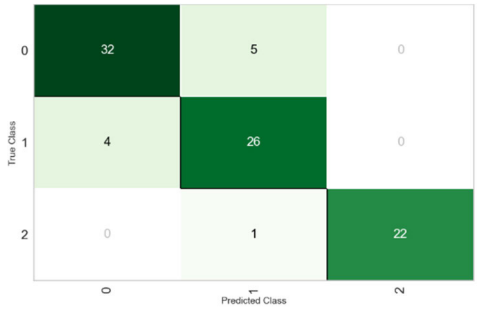


FIGURE 18. Confusion matrix obtained for "Millennials."

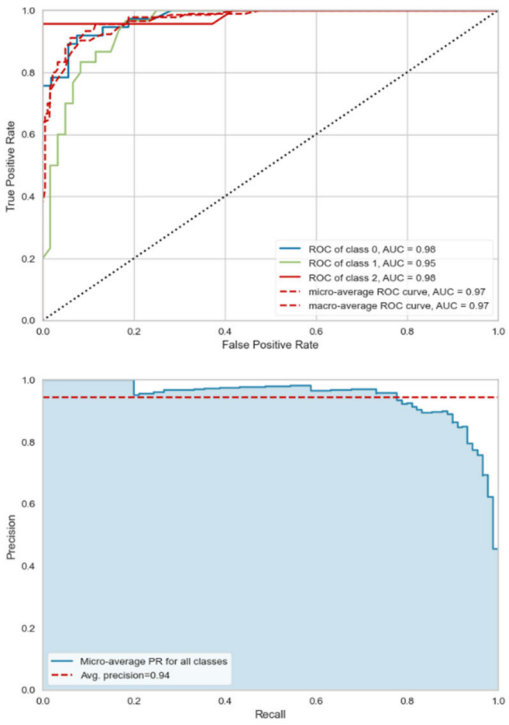


FIGURE 19. AUC and Precision-Recall curves obtained for "Millennials."

a) Confusion Matrix

It is evident from Figure 18 that the proposed ensemble model performs best in the case of predicting diabetes (2), followed by non-diabetic (0) and pre-diabetic (1) respectively.

b) AUC and Precision and Recall

It can be seen from Figure 19 the classifier is most accurate in predicting class 0 and 2, with an AUC value of 0.98. This means that the classifier correctly predicts the class of 98% of all samples from class 0 and 2. The classifier is also very accurate in predicting class 1, with an AUC value of 0.95. This means that the classifier correctly predicts the class of 95% of all samples from class 1. The classifier is able to avoid false positives 94% of the time from PR graph.

c) Class Prediction Error

From the above graph in Figure 20, it can clearly be seen that the model classifies diabetic (2) people most accurately as it has the least prediction error, followed by non-diabetic (0) and lastly pre-diabetic (1).

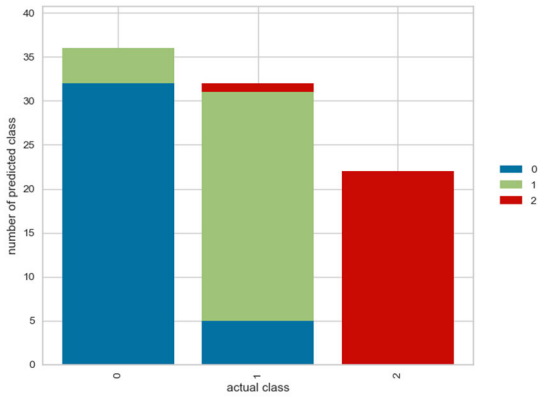


FIGURE 20. Class prediction error graph obtained for "Millennials" where 0:non-diabetic, 1: pre-diabetic and 2: diabetic.

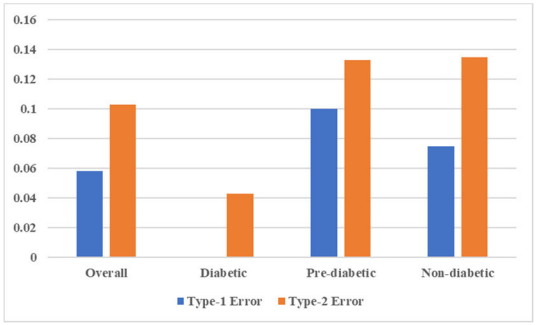


FIGURE 21. Type-1 and Type-2 error graph obtained for "Millennials."

FIGURE 22. Confusion matrix obtained for "30-40 year age group."

TABLE 12. Type-1 and Type-2 error values obtained for "Millennials."

Millennials (18-30)	Type-1 Error	Type-2 Error
Overall	0.058	0.103
Diabetic	0	0.043
Pre-diabetic	0.1	0.133
Non-diabetic	0.075	0.135

d) Type-1 and Type-2 Error

The TABLE 12 and Figure 21 presents the Type-1 and Type-2 error rates for different age groups, specifically focusing on millennials aged 18 to 30. Among millennials, the Type-1 error rate is recorded as 0.058, indicating a moderate occurrence of false positive diagnoses. The Type-2 error rate is slightly higher at 0.103, suggesting a moderate likelihood

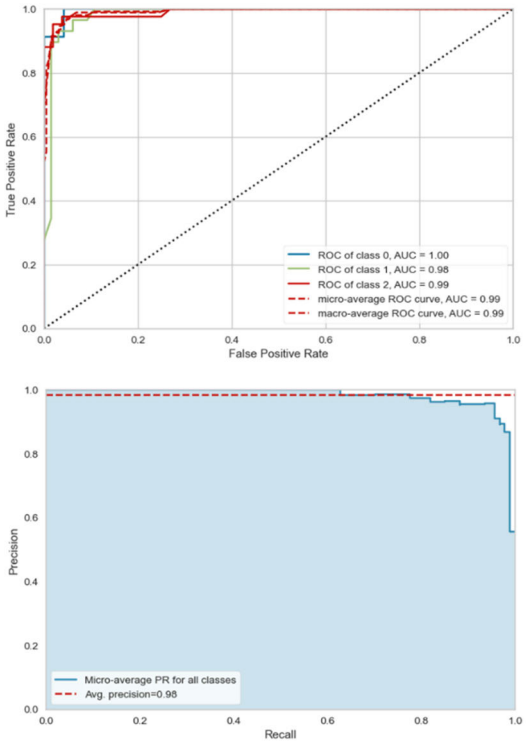


FIGURE 23. AUC and Precision-Recall curves obtained for "30-40 year age group."

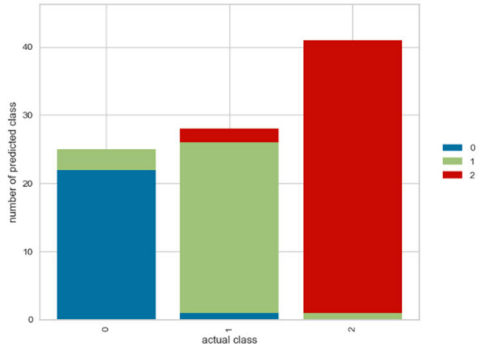


FIGURE 24. Class Prediction error graph obtained for "30-40 year age group" where 0:non-diabetic, 1: pre-diabetic and 2: diabetic.

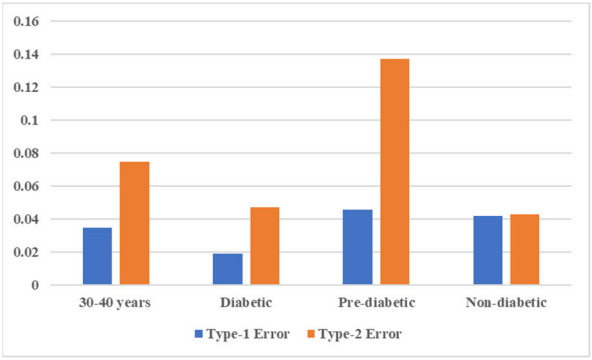


FIGURE 25. Type-1 and Type-2 error graph obtained for "30-40 year age group."

of false negative diagnoses. For millennials identified as diabetic, the Type-1 error rate is 0, indicating an absence

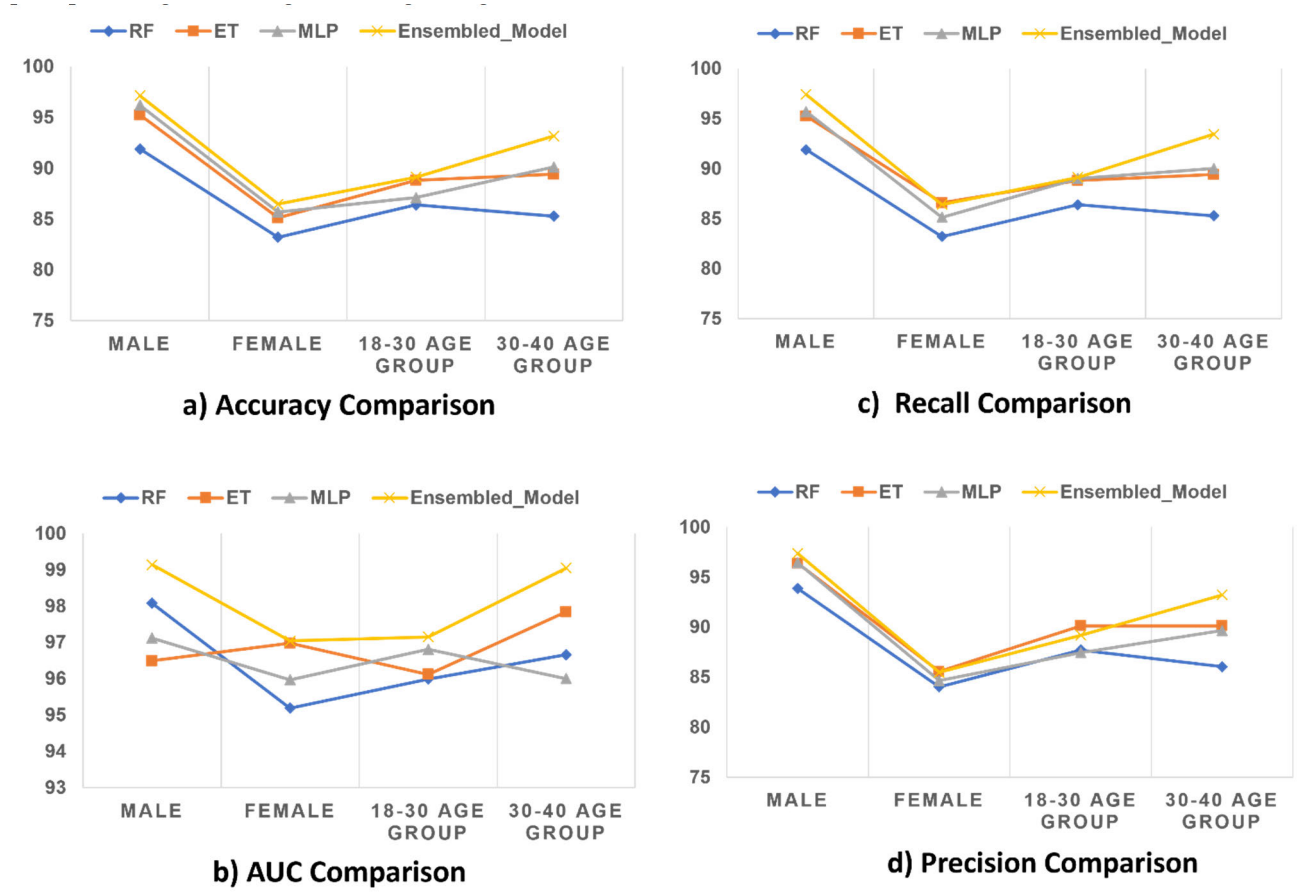


FIGURE 26. Comparison of proposed ensembled model and based models.

TABLE 13. Evaluation metrices values obtained for "30-40 Year Age Group."

Evaluation Metrics	Value
Accuracy	0.931
AUC	0.990
Recall	0.934
Precision	0.932
F1	0.933

of false positive diagnoses. However, the Type-2 error rate is 0.043, reflecting a relatively low chance of false negative diagnoses. Among millennials classified as pre-diabetic, the Type-1 error rate is notably higher at 0.1, signifying a significant probability of false positive diagnoses, while the Type-2 error rate stands at 0.133, suggesting a moderate likelihood of false negative diagnoses. Notably, for millennials identified as non-diabetic, the Type-1 error rate is 0.075, indicating a moderate occurrence of false positive diagnoses, while the Type-2 error rate is recorded as 0.135, suggesting a moderate probability of false negative diagnoses.

E. AGE INTERVAL ANALYSIS (AGE GROUP (30-40))

The TABLE 13 depicts the evaluation metrics value obtained for age group 30-40 after applying the proposed ensembled model.

a) Confusion Matrix

It is evident from the Figure 22 that the proposed ensemble model performs best in the case of predicting diabetic (2), followed by non-diabetic (0) and pre-diabetic (1) respectively.

b) AUC and Precision and Recall

It can be seen from Figure 23 the classifier is most accurate in predicting class 0, with an AUC value of 1.00. This means that the classifier correctly predicts the class of 100% of all samples from class 0. The classifier is slightly less accurate in predicting class 1, with an AUC value of 0.98. This means that the classifier correctly predicts the class of 98% of all samples from class 1. The classifier is also very accurate in predicting class 2, with an AUC value of 0.99. This means that the classifier correctly predicts the class of 99% of all samples from class 2. The classifier is able to avoid false positives 98% of the time from PR graph.

c) Class Prediction Error

From the graph in Figure 24, it can clearly be seen that the model classifies diabetic (2) people most accurately as it has the least prediction error, followed by non-diabetic (0) and lastly pre-diabetic (1).

d) Type-1 and Type-2 Error

The provided TABLE 14 and Figure 25, presents the Type-1 and Type-2 error rates for individuals in the age

TABLE 14. Type-1 and Type-2 error values obtained for "30-40 Year Age Group."

30-40 Age Group	Type-1 Error	Type-2 Error
Overall	0.035	0.075
Diabetic	0.019	0.047
Pre-diabetic	0.046	0.137
Non-diabetic	0.042	0.043

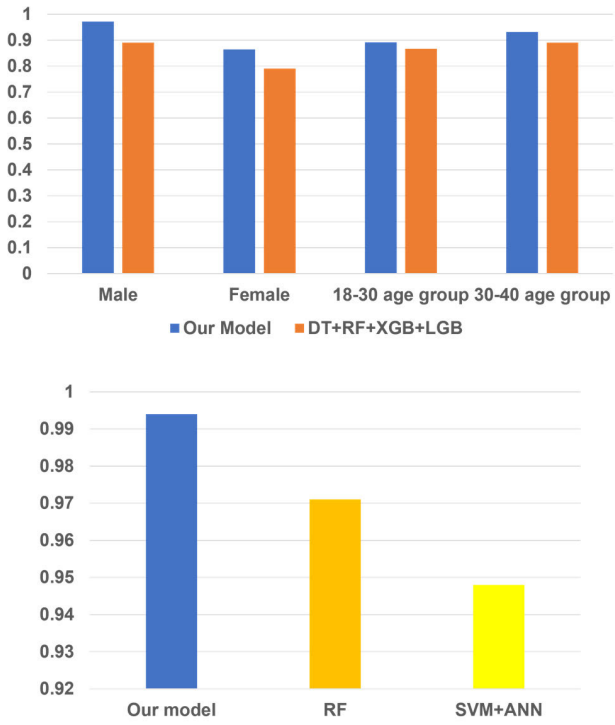


FIGURE 27. Comparison with other related works.

group of 30 to 40 years. Among individuals aged 30 to 40, the Type-1 error rate is recorded as 0.035, indicating a relatively low occurrence of false positive diagnoses. The Type-2 error rate is slightly higher at 0.075, suggesting a moderate likelihood of false negative diagnoses. For individuals in this age group identified as diabetic, the Type-1 error rate is 0.019, reflecting a low probability of false positive diagnoses. Similarly, the Type-2 error rate is 0.047, indicating a relatively low chance of false negative diagnoses. Among individuals classified as pre-diabetic in the 30 to 40 years age group, the Type-1 error rate is 0.046, signifying a moderate probability of false positive diagnoses, while the Type-2 error rate stands at 0.137, suggesting a higher likelihood of false negative diagnoses. Notably, for individuals identified as non-diabetic in this age group, the Type-1 error rate is 0.042, indicating a relatively low occurrence of false positive diagnoses, while the Type-2 error rate is 0.043, reflecting a similar likelihood of false negative diagnoses.

TABLE 15. Comparison with related works.

	Classifier used	Dataset	Dataset Size	Model designed	Accuracy
[19]	Weighted Classifier	Newly labelled Diabetes dataset from Bangladesh	12,119	DT+RF+XGB+LGB	0.79 (Male) 0.89 (female) 0.88 (millennials) 0.78 (30-40 age group)
[18]	Hard Voting Classifier	UCI Repository	520	SVM+ANN	0.95 (UCI dataset)
[39]	Random Forest Ensemble Classifier	UCI Repository	520	RF	0.97 (UCI dataset)
Our Model	Soft Voting Classifier	Diabetes Dataset from Care Hospital and UCI Repository	672	RT+ET+MLP	0.994 (UCI dataset) 0.864 (female) 0.971 (male) 0.891 (millennials) 0.931 (30-40 age group)

F. COMPARISON WITH BASE MODELS

From the Figure 26, line graph, it's evident that the ensembled model consistently outperforms all other base models in terms of accuracy, recall, precision and AUC, including random forest, extra tree, and multi-layer perceptron, regardless of gender or age categories.

G. COMPARISON WITH OTHER RELATED WORKS

To demonstrate the better performance of our proposed model compared to models introduced by the authors in [19], we applied their model to our dataset. The outcomes revealed that, across all age and gender categories, our model exhibited superior performance shown in Figure 27 and TABLE 15. Additionally, when contrasting our proposed model with the study conducted by authors in [18] and [39], where the UCI repository dataset was used we applied our proposed ensemble model to it, the results have shown that our accuracy turned out to be better when compared to others showcasing the generalizability of our proposed model.

IX. CONCLUSION AND DISCUSSION

In conclusion, to acquire a thorough grasp of the wide-ranging impact and implications of the condition, this study examined the influence of gender and age variations in people with diabetes. A real-world dataset obtained from a diabetologist, maintained by Dr. Reddy's Lab, was utilized for analysis. The effectiveness of the predictive model was improved by using an ensemble learning technique, producing precise predictions of the presence of diabetes. The approach offers a ternary label that designates whether a person is considered to have diabetes, not to have diabetes, or to

have pre-diabetes. In addition, the model gives a prediction score that expresses how likely it is that a person falls into each group.

Important findings and conclusions:

- The analysis revealed that certain features were found to be important for both males and females in predicting diabetes, such as fasting glucose and HbA1c.
- However, gender-specific variations were also observed. In males, uric acid levels and systolic blood pressure emerged as important contributing factors, highlighting the potential influence of cardiovascular health. Conversely, for females, total cholesterol and triglyceride levels played a prominent role, emphasizing the importance of lipid metabolism in diabetes prediction.
- Fasting glucose and HbA1c levels are important predictors of diabetes in the millennial and 30–40-year age group. Gender has a notable influence on diabetes prediction among millennials and adult group.
- Triglyceride levels, body mass index (BMI), and total cholesterol also significantly contribute to diabetes prediction in this population, whereas total cholesterol and diastolic blood pressure are important factors to consider for accurate diabetes prediction in this age range.
- Notably, the proposed ensembled model's predictive capability is particularly noteworthy for males, while its accuracy is higher for adults in the age range of 30 to 40 years.
- Irrespective of age or gender, the proposed ensembled model exhibits the highest accuracy in predicting diabetic patients, followed by non-diabetic and pre-diabetic individuals.
- In all the categories of observation (gender and age) the proposed ensembled model outperforms the rest of the base models (RF, MLP, ET).
- Dr. K.D Modi mentioned that the proposed model categories overall risk status by age and gender, offering valuable insights for doctors to focus on specific patients for future cardiovascular and metabolic risk evaluation.

These observations emphasize the importance of considering gender and age factors when employing the ensemble model for predicting diabetes. One potential avenue for future research involves expanding the demographic considerations to encompass a broader spectrum of age groups and ethnicities. The practical applications of this research extend into the realm of healthcare decision support systems, where the developed models could be integrated to assist clinicians in diagnosing and managing diabetes more effectively.

One notable limitation of our study is the relatively small sample size, comprising only 672 patients, which may restrict the generalizability of our proposed algorithm. Moreover, the dataset originates solely from a single hospital (Care hospital, Dr. Modi) in Hyderabad, India, potentially introducing bias towards a specific demographic. Consequently, future research endeavours would greatly benefit from applying our algorithm to a larger and more diverse population, encompassing datasets from various regions across the country. This approach would enable a more comprehensive evaluation of the algorithm's performance and enhance its applicability in broader clinical settings. In addition, the inclusion of additional relevant features, such as lifestyle factors, socioeconomic variables, or genetic markers, can enhance the predictive power of the model. An age group other greater than 40 years can also be explored. By pursuing these future directions, we can improve disease management and healthcare outcomes for individuals at risk of diabetes.

ACKNOWLEDGMENT

Rishi Jain deeply indebted to Dr. K. D. Modi, a distinguished Consultant in Diabetologist and Endocrinology with CARE Hospitals, Nampally, Hyderabad. The valuable dataset used in this article was graciously provided by Dr. Modi. Dr. K. D. Modi holds an impressive academic background, with qualifications including MD (Internal Medicine), DM (Endocrinology), and DNB (Endocrinology). His expertise in the field of endocrinology and his commitment to advancing medical knowledge have significantly enriched the quality of this research.

REFERENCES

[1] U. Ahmed, G. F. Issa, M. A. Khan, S. Aftab, M. F. Khan, R. A. T. Said, T. M. Ghazal, and M. Ahmad, "Prediction of diabetes empowered with fused machine learning," *IEEE Access*, vol. 10, pp. 8529–8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

[2] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Exp.*, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

[3] 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, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[4] M. Sevil, M. Rashid, I. Hajizadeh, M. Park, L. Quinn, and A. Cinar, "Physical activity and psychological stress detection and assessment of their effects on glucose concentration predictions in diabetes management," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 7, pp. 2251–2260, Jul. 2021, doi: 10.1109/TBME.2020.3049109.

[5] A. Issaka, A. J. Cameron, Y. Paradies, W. K. Bosu, Y. C. N. Houehanou, J. B. Kiwallo, C. S. Wesseh, D. S. Houinato, D. J. P. Nazoum, and C. Stevenson, "Effect of age and sex on the associations between potential modifiable risk factors and both type 2 diabetes and impaired fasting glycaemia among west African adults," *BMC Public Health*, vol. 22, no. 1, pp. 1–12, Jun. 2022, doi: 10.1186/s12889-022-13588-w.

[6] L.-C. Huang, C.-L. Lin, Y.-T. Chang, R.-Y. Chen, and C.-H. Bai, "Gender impact on diabetes distress focus at medical communication concerns, life and interpersonal stress," *Int. J. Environ. Res. Public Health*, vol. 19, no. 23, p. 15678, Nov. 2022, doi: 10.3390/ijerph192315678.

[7] C. Deischinger, E. Dervic, M. Leutner, L. Kosi-Trebotic, P. Klimek, A. Kautzky, and A. Kautzky-Willer, "Diabetes mellitus is associated with a higher risk for major depressive disorder in women than in men," *BMJ Open Diabetes Res. Care*, vol. 8, no. 1, Sep. 2020, Art. no. e001430, doi: 10.1136/bmjdrc-2020-001430.

[8] L. Lama, O. Wilhelmsson, E. Norlander, L. Gustafsson, A. Lager, P. Tynelius, L. Wärvik, and C.-G. Östenson, "Machine learning for prediction of diabetes risk in middle-aged Swedish people," *Heliyon*, vol. 7, no. 7, Jul. 2021, Art. no. e07419, doi: 10.1016/j.heliyon.2021.e07419.

[9] Y. Tian, C. Li, T. A. Shilko, V. S. Sosunovsky, and Y. Zhang, "The relationship between physical activity and diabetes in middle-aged and elderly people," *Medicine*, vol. 102, no. 6, 2023, Art. no. e32796, doi: 10.1097/md.0000000000032796.

[10] M. Xue, X. Yang, Y. Zou, T. Liu, Y. Su, C. Li, H. Yao, and S. Wang, "A non-invasive prediction model for non-alcoholic fatty liver disease in adults with type 2 diabetes based on the population of Northern Urumqi, China," *Diabetes, Metabolic Syndrome Obesity, Targets Therapy*, vol. 14, pp. 443–454, Feb. 2021, doi: 10.2147/dmso.s271882.

[11] P. Kakria, N. K. Tripathi, and P. Kitipawang, "A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors," *Int. J. Telemedicine Appl.*, vol. 2015, pp. 1–11, Jan. 2015, doi: 10.1155/2015/373474.

[12] W. Zheng, J. Chu, J. Ren, J. Dong, H. Bambrick, N. Wang, K. Mengersen, X. Guo, and W. Hu, "Age- and gender-specific differences in the seasonal distribution of diabetes mortality in Shandong, China: A spatial analysis," *Int. J. Environ. Res. Public Health*, vol. 19, no. 24, p. 17024, Dec. 2022, doi: 10.3390/ijerph192417024.

[13] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[14] (2023). *Dr. Reddy's Laboratories—Good Health Can't Wait*. Accessed: Dec. 26, 2023. [Online]. Available: https://www.drreddys.com/

[15] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, nos. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039.

[16] Y. Su, C. Huang, W. Yin, X. Lyu, L. Ma, and Z. Tao, "Diabetes mellitus risk prediction using age adaptation models," *Biomed. Signal Process. Control*, vol. 80, Feb. 2023, Art. no. 104381, doi: 10.1016/j.bspc.2022.104381.

[17] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, p. 3317, Mar. 2021, doi: 10.3390/ijerph18063317.

[18] T. R. Mahesh, D. Kumar, V. V. Kumar, J. Asghar, B. M. Bazezew, R. Natarajan, and V. Vivek, "Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–9, Jul. 2022, doi: 10.1155/2022/4451792.

[19] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, and H. Meshref, "Early prediction of diabetes using an ensemble of machine learning models," *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, p. 12378, Sep. 2022, doi: 10.3390/ijerph191912378.

[20] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: A retrospective cohort study," *J. Diabetes Metabolic Disorders*, vol. 21, no. 1, pp. 251–261, Jan. 2022, doi: 10.1007/s40200-021-00968-z.

[21] A. Chen, W. Zhou, J. Hou, A. Nevill, Y. Ding, Y. Wan, R. Jester, X. Qin, Z. Hu, and R. Chen, "Impact of older age adiposity on incident diabetes: A community-based cohort study in China," *Diabetes Metabolism J.*, vol. 46, no. 5, pp. 733–746, Sep. 2022, doi: 10.4093/dmj.2021.0215.

[22] S. K. Kunutsor, A. Voutilainen, and J. A. Laukkanen, "Handgrip strength improves prediction of type 2 diabetes: A prospective cohort study," *Ann. Med.*, vol. 52, no. 8, pp. 471–478, Nov. 2020, doi: 10.1080/07853890.2020.1815078.

[23] Y. Zhang, M. Tong, B. Wang, Z. Shi, P. Wang, L. Li, Y. Ning, and T. Lu, "Geographic, gender, and seasonal variation of diabetes: A nationwide study with 1.4 million participants," *J. Clin. Endocrinol. Metabolism*, vol. 106, pp. 4981–4992, Jul. 2021, doi: 10.1210/clinem/dgab543.

[24] X. T. Cai, L. W. Ji, S. S. Liu, M. R. Wang, M. Heizhati, and N. F. Li, "Derivation and validation of a prediction model for predicting the 5-year incidence of type 2 diabetes in non-obese adults: A population-based cohort study," *Diabetes, Metab. Syndr. Obes.*, vol. 14, pp. 2087–2101, May 2021, doi: 10.2147/DMSO.S304994.

[25] (2024). *Diabetes Tests | CDC*. Accessed: Mar. 11, 2024. [Online]. Available: https://www.cdc.gov/diabetes/basics/getting-tested.html

[26] 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, no. 1, pp. 1–14, Dec. 2020, doi: 10.1007/s13755-019-0095-z.

[27] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M. H. Rahman, "Prevalence and early prediction of diabetes using machine learning in North Kashmir: A case study of district bandipora," *Comput. Intell. Neurosci.*, vol. 2022, 2022, Art. no. 2789760, doi: 10.1155/2022/2789760.

[28] L. Chang, Y. Fukuoka, B. E. Aouizerat, L. Zhang, and E. Flowers, "Prediction performance of feature selectors and classifiers on highly dimensional transcriptomic data for prediction of weight loss in Filipino Americans at risk for type 2 diabetes," *Biol. Res. Nursing*, vol. 25, no. 3, pp. 393–403, Jan. 2023, doi: 10.1177/10998004221147513

[29] S. Mishra, H. K. Tripathy, P. K. Mallick, A. K. Bhoi, and P. Barsocchi, "EAGA-MLP—An enhanced and adaptive hybrid classification model for diabetes diagnosis," *Sensors*, vol. 20, no. 14, p.4036, 2020.

[30] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," *J. Healthcare Eng.*, vol. 2021, pp. 1–17, Sep. 2021, doi: 10.1155/2021/9930985.

[31] S. K. Kalagotla, S. V. Gangashetty, and K. Giridhar, "A novel stacking technique for prediction of diabetes," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104554, doi: 10.1016/j.compbiomed.2021.104554.

[32] A. Dogru, S. Buyrukoglu, and M. Ari, "A hybrid super ensemble learning model for the early-stage prediction of diabetes risk," *Med. Biol. Eng. Comput.*, vol. 61, no. 3, pp. 785–797, Mar. 2023, doi: 10.1007/s11517-022-02749-z.

[33] M. J. Sai, P. Chettri, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, "An ensemble of light gradient boosting machine and adaptive boosting for prediction of type-2 diabetes," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 14, Feb. 2023, doi: 10.1007/s44196-023-00184-y.

[34] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," *IEEE Access*, vol. 9, pp. 103737–103757, 2021, doi: 10.1109/ACCESS.2021.3098691.

[35] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cognit. Comput. Eng.*, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[36] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, Art. no. 106773, doi: 10.1016/j.cmpb.2022.106773.

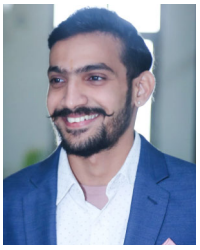
[37] A. Singh, A. Dhillon, N. Kumar, M. S. Hossain, G. Muhammad, and M. Kumar, "EDiaPredict: An ensemble-based framework for diabetes prediction," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 2, pp. 1–26, Jun. 2021, doi: 10.1145/3415155.

[38] J. Xue, F. Min, and F. Ma, "Research on diabetes prediction method based on machine learning," *J. Phys., Conf.*, vol. 1684, Nov. 2020, Art. no. 012062, doi: 10.1088/1742-6596/1684/1/012062.

[39] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study," *Sensors*, vol. 22, no. 14, p. 5247, Jul. 2022, doi: 10.3390/s22145247.

[40] S. Gündogdu, "Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique," *Multimedia Tools Appl.*, vol. 82, no. 22, pp. 34163–34181, Sep. 2023, doi: 10.1007/s11042-023-15165-8.

[41] *Diabetes UCI Dataset*. Accessed: Dec. 27, 2023. [Online]. Available: https://www.kaggle.com/datasets/alakaaay/diabetes-uci-dataset



RISHI JAIN received the B.Tech. degree (Hons.) in computer science from Jawaharlal Nehru Technology University, Hyderabad, Telangana, India, in 2020, and the M.E. degree (Hons.) in information management from Asian Institute of Technology, Thailand, in 2021.

He was a Research Assistant and a Teaching Assistant with Maastricht University, The Netherlands, under the guidance of Dr. Clemens Betcher. He was also a Project Assistant with Asian Institute of Technology, under the guidance of Prof. Nitin Kumar Tripathi. His research interests include machine learning and its applications, health informatics, and artificial intelligence. He is currently a dual doctoral Ph.D. Scholar with Asian Institute of Technology, Thailand, and Indian Institute of Technology Roorkee, India. In addition, he is also a Trainee under a machine learning project at Deloitte.

Mr. Jain received a fully funded and partially funded scholarship from Indian Institute of Technology Roorkee, India, and Asian Institute of Technology, Thailand, respectively. He is also serving as a Reviewer for the *Engineering Applications of Artificial Intelligence journal* and for the Soft Computing and Problem-Solving Conference. He has two accepted papers in 12th International Conference SocProS 2023 and was invited to INCOST 2023 Conference as a Speaker.

NITIN KUMAR TRIPATHI received the B.Tech. degree in civil engineering from NIT Warangal, India, in 1984, and the M.Tech. degree in remote sensing and the Ph.D. degree in remote sensing from IIT Kanpur.

He has a total of 224 publications to his credit (three book, 12 chapters in books, 145 research articles in peer-reviewed journals, and 65 conference papers) mostly in peer-reviewed international journals. His Scopus H-index is 30 with 3337 citations and in Google Scholar H-index 37 with 5770 citations. He was with the National Institute of Technology, Allahabad, India (1988–1989) and Indian Institute of Technology (IIT) Kanpur, India (1989–1999), before joining AIT, in 2000. He is also a Professor in remote sensing (RS) and geographical information systems (GIS) with Asian Institute of Technology (AIT), Thailand. His research interests include the applications of GIS and RS in climate change impacts on water resources, agriculture and health, health and environment, radio-frequency identification (RFID) and location-based information services, internet GIS and system security, and crop substitution modeling.

Prof. Tripathi was actively involved as a Life Member of Indian Society of Remote Sensing and Indian Society of Geomatics. He also served as a Life Member for Institution of Engineers—India and also Indian Society of Technical Education-India. He was an active member of Asian Conference on Remote Sensing and launched the first *Asian Journal of Geoinformatics* from AIT, in 2001. Later, he launched another *International Journal of Geoinformatics*, in 2005, and is the Editor-in-Chief. He is in the editorial board of *Journal of Remote Sensing* (Springer) and ISRS (Taylor and Francis). He has been serving on the advisory board of a number of international journals. He has earned several awards, including Young Scientist Award from the Department of Atomic Energy, India, AICTE Career Award for Young Teacher, MHRD-India, Osaka City University Distinguished Scientist Award, Japan, a Navigator of Thailand by Royal Thai Survey Department, and several others including best paper awards in conference papers. In August 2022, he received Excellence Award—Education from Indo_Thai TV in Thailand. He has been serving on the advisory board of a number for international journals.

MILLIE PANT received the M.Sc. degree in mathematics from CCS University, Meerut, and the Ph.D. degree from the Department of Mathematics, IIT Roorkee, India. She is currently a Professor and the Head of the Department of Applied Mathematics and Scientific Computing, IIT Roorkee, where she is also a Joint Faculty Member of the Mehta Family School of Data Science and Artificial Intelligence. She has authored or coauthored more than 350 papers in various journals and conferences of national and international repute. Her research interests include numerical optimization, operations research, evolutionary algorithms, supply chain management, swarm intelligence techniques, AI-assisted decision making, data analysis, and image processing. She is a Guest Reviewer of IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, *Applied Mathematics and Computation*, and *Applied Soft Computing*. She is also an Associate Editor of the *International Journal of Swarm Intelligence* (Inderscience) and a Guest Editor of the *International Journal of Memetic Computing* (Springer).

CHUTIPORN ANUTARIYA received the B.Sc. degree (Hons.) in statistics from Chulalongkorn University, Thailand, and the M.Sc. and D.Tech.Sc. degree in computer science from Asian Institute of Technology, Thailand.

She is currently an Associate Professor and also the Head of the Department of ICT, Asian Institute of Technology. Her research interests include learning technologies and massive open online courses (MOOCs), database and information systems, knowledge representation and knowledge management, open data and open government data, semantic and linked data technologies, ontologies, and web and mobile technologies. She has received several awards and honors namely, Royal Golden Jubilee Ph.D. Program Grant Awardee, Thailand Research Fund (TRF), in 2007, Best Research Award (2007–2008) and (2005–2006), Shinawatra University, Dissertation Award, in 2003, National Research Council of Thailand (NRCT), Gold Medal Prize, in 1996, and Chulalongkorn University.

CHAKLAM SILPASUWANCHAI received the Bachelor of Science degree (Hons.) in computer science from the Sirindhorn International Institute of Technology, Thailand, the Master of Engineering degree in computer science from Asian Institute of Technology, Thailand, and the Doctor of Engineering degree in computer science from the Kochi University of Technology, Japan.

From 2015 to 2017, he was a Postdoctoral Researcher with the Kochi University of Technology. He was a Faculty Member of the IT Program, Stamford International University, Thailand (2017–2019). Since 2019, he has been an Assistant Professor with the Computer Science and Information Management Department, Asian Institute of Technology. His research interests include machine/deep learning, natural language processing, software engineering, hypothesis testing, and data structures and algorithms.