RESEARCH ARTICLE OPEN ACCESS

Manuscript received July 2, 2024; revised August 12, 2024; accepted August 16, 2024; date of publication August 20, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i4.497

How to cite: Rohini Patil , Anant Patil , Surekha Janrao , Sandip Bankar, and Kamal Shah, "A Framework for Prediction of Type II Diabetes through Ensemble Stacking Model", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 4, pp. 459-466, October 2024.

A Framework for Prediction of Type II Diabetes through Ensemble Stacking Model

Rohini Patil[1], Anant Patil[2], Surekha Janrao[3], Sandip Bankar[4], and Kamal Shah[5]

[1]Department of Computer Engineering, Terna Engineering College, Navi Mumbai, India
[2]Department of Pharmacology, DY Patil Deemed to be University School of Medicine, Navi Mumbai, India
[3]Department of Computer Engineering, K. J. Somaiya Institute of Technology, Mumbai, India
[4]NMIMS University, Navi Mumbai, India
[5]St John College of Engineering and Management, Palghar, India

ABSTRACT In order to prevent long term complications of diabetes its early diagnosis is crucial. With Increasing advances in Artifical Intelligence (AI) and Machine Learning(ML) researchers are increasingly focusing on using them for early diagnosis of diseases.AI and ML has significant potential for early prediction of type 2 diabetes. This study aims to demonstrate results of a ML based framework for early prediction of type 2 diabetes -Improved Ensemble Learning with Dimensionality Reduction Model (IELDR). Implementation and assessment of the proposed IELDR algorithm predicts the possibility of developing type 2 diabetes by utilizing a comprehensive self-collected LS_diabetes dataset. The effectiveness of the developed models based on prediction accuracy has been validated through a thorough comparison against cutting-edge methods. An IELDR algorithm is an Auto encoder-based feature extraction method with ensemble learning. LS_diabetes dataset containing 374 records with 35 features related to lifestyle and stress. Accuracy, precision, specificity, sensitivity, f1 score, roc and Mathew correlation coefficient (MCC) were measured. After this results were tested and validated using Diabetes_2019 dataset and PIMA diabetes dataset. The IELDR model showed results in terms accuracy, precision, specificity, sensitivity, f1 score, roc and Mathew correlation coefficient (MCC) of 98.67%, 95.24%, 100%, 98.18%, 97.56%, 99.09% and 0.97 respectively. In comparison with PIMA diabetes dataset, LS_diabetes dataset showed an accuracy, precision, sensitivity, specificity, f1-score,roc and mcc value by 17.96%,13.15% 40.22%,5.59%,28.38%,22.09% and 0.4 respectively. The IELDR model achieved the best result on the LS_diabetes dataset showed an accuracy, sensitivity, roc and mcc value improved by 1.82%, 1.58%, 3.01%and 0.04 % compared to the Diabetes_2019 dataset .This proposed IELDR system predicts the risk of type 2 diabetes in a healthy person based on the person's current lifestyle pattern. This system can be helpful for early prediction of type2 diabetes.

INDEX TERMS: Machine Learning, Diabetes, Prediction, Ensemble Learning, Risk, Lifestyle, Stress.

I. INTRODUCTION

Lifestyle diseases are driven by several factors, including the globalization, rapid urbanization, and unhealthy habits such as poor nutrition, sleep deprivation, stress, and a sedentary work. Lifestyle diseases are preventable, and their occurrence can be reduced through dietary and lifestyle changes. Early prediction of such diseases is important to improve the quality of life and decrease risk of complications. Machine learning (ML) is used in many domains, including the medical field [1-3]. Many models are designed for type 2 diabetes mellitus [4-6], depression [7-8], asthma [9], stroke [10], metabolic syndrome [11], heart disease [12], osteoporosis, acne and obesity.

Diabetes is one of the most challenging diseases to manage from a psychosocial and behavioral standpoint. According to the WHO report (2019), diabetes is the 9th cause of death [13]. In 2019, it was anticipated that there would be 9.3 % of diabetes worldwide by 2045, increasing from 10.2% to 10.9% [14]. Global Report published in 2017[14] by the International Diabetes Federation claims that there are 82 million adults with diabetes. South East Asia is the second highest region among all other areas. India contributes about 49% of the world's burden. In Southeast Asia, out of 88 million people with diabetes, India contributes 77 million, which is expected to increase to 134.2 million in 2045[14].

Effective diabetes management includes early detection, preventing short- and long-term morbidity, and promoting self-care practices [15]. To reduce the incidence of disease, preventive measures such as avoiding smoking and being overweight, regular physical activity, consumption of healthy types of fat, eating plenty of fruits and vegetables, replacing refined grains with whole grains, avoiding excessive calories, etc., are necessary. Physical activity or exercise increases metabolic control, insulin sensitivity, and cardiorespiratory fitness and helps to maintain body weight. Diabetes mellitus is a global issue with research challenges across the world and in India [13-15]

The intention of using artificial intelligence (AI) and ML in healthcare is to increase the diagnostic accuracy and effectiveness of therapy and help clinicians in their practice of patient management with improved outcomes. The decision to use AI in the diagnostic field is quickly gaining momentum due to improvements in accuracy and a massive amount of data availability [16]. Early diagnosis and better care at a reasonable cost are critical to improving patient satisfaction. Enormous data and increased complexities have led to rising interest in using machine learning in healthcare. The growth of healthcare coupled with technology has fueled the expanded need for ML in healthcare. As expected, ML helps to gain important information from enormous amounts of available data. ML is helping healthcare by providing innovative and relevant information, which is practically difficult to analyse manually, given time constraints, human resources, and other resources [17]. In recent years, many researchers have used ML algorithms to predict diabetes mellitus [18-22]. The high prevalence and associated complications of diabetes underline the importance of early detection of the disease and the need for prompt measures to control it. Early detection is possible by regular screening of glucose level and glycosylated haemoglobin levels. However, it requires high motivation of the person for screening at regular intervals. Considering this background, there is an unmet need for a cost-efficient, convenient and accurate ML algorithm that can assist in the early prediction. Most of the studies have worked on PIMA Indian diabetes dataset, which is female-centric [18-20,22]. Limited data exists on prediction of type 2 diabetes based on stress and lifestyle factors [18]. There is significant scope of performance improvement of various ML algorithms for early diagnosis of diabetes and it is worthy of further research. The contributions of this work are as follows:

1. The study aimed for early diagnosis of type 2 diabetes disease using the proposed ensemble method (IELDR). The proposed stacking model was applied to solve classification using MLP, KNN, LR, SVM, GBC, RF, LDA, DT.
2. The key contribution is the data generated from the region in Maharashtra to find the stress and lifestyle effect on diabetes disease.
3. The ensemble method aims to improve the accuracy and performance of the model for complex, noisy, and imbalanced data.
4. The study also aimed to validate the results using various parameters such as accuracy, precision, specificity, sensitivity, f1 score, ROC, and MCC.

The study is organized as follows: The study methodology and a description of the datasets are both discussed in section 2. Section 3 concentrates on results and section 4 discussion, and the final section 5 presents the conclusion.

II. METHODOLOGY

A practical model is required to detect diabetes more accurately as early as possible. ML is needed for automation with minimum human effort. A single algorithm approach to predict diabetes would not provide a reliable and efficient prediction system. To increase the accuracy of innovative prediction, we proposed a novel auto encoder-based ensemble framework –IELDR. The IELDR model was employed on the Lifestyle stress diabetes dataset (LS_diabetes dataset), PIMA diabetes dataset [23] , and Diabetes_2019 dataset [24] to obtain a classification.

A. DATASETS

The proposed novel framework used LS_diabetes dataset which contains total of 374 records with 35 features of

TABLE 1
Dataset description

Dataset	Samples		Attributes
Lifestyle stress diabetes dataset	Diabetes	86	Gender, Age, Height, Weight, BMI, Anxiety, Stress, Workload, Satisfaction, Profession, profile, Smoke, Exercise, Cereal grains consumption, Salad, Cooked Vegetables, Sweets, Sweet_freq, Sugar, Milk Consumption, Milk quantity consumption, High blood pressure, Systolic, Diastolic, Fasting Sugar, Post meal sugar, HbA1c, Family history, diabetes
	Non_Diabetes	288	
	Total_Records	374	
	Input attributes	34	
	Total_Atributes	35	
Diabetes_2019 dataset [23]	Diabetes	267	Age, Gender, Family history, High blood pressure, Walk/run/physically active, BMI, Smoking, Alcohol consumption, Sleep, Sound sleep, Daily medicine intake, Junk food intake, Stress, Blood pressure level, Number of pregnancies, Gestation diabetes, Urination frequency,diabetic
	Non_Diabetes	685	
	Total_Records	952	
	Input attributes	17	
	Total_Atributes	18	
PIMA diabetes dataset [24]	Diabetes	268	Preg, Plas, Test, DPres, Skin_T, Insul, BMI, Pedig, Age,class
	Non_Diabetes	500	
	Total_Records	768	
	Input attributes	08	
	Total_Atributes	09	

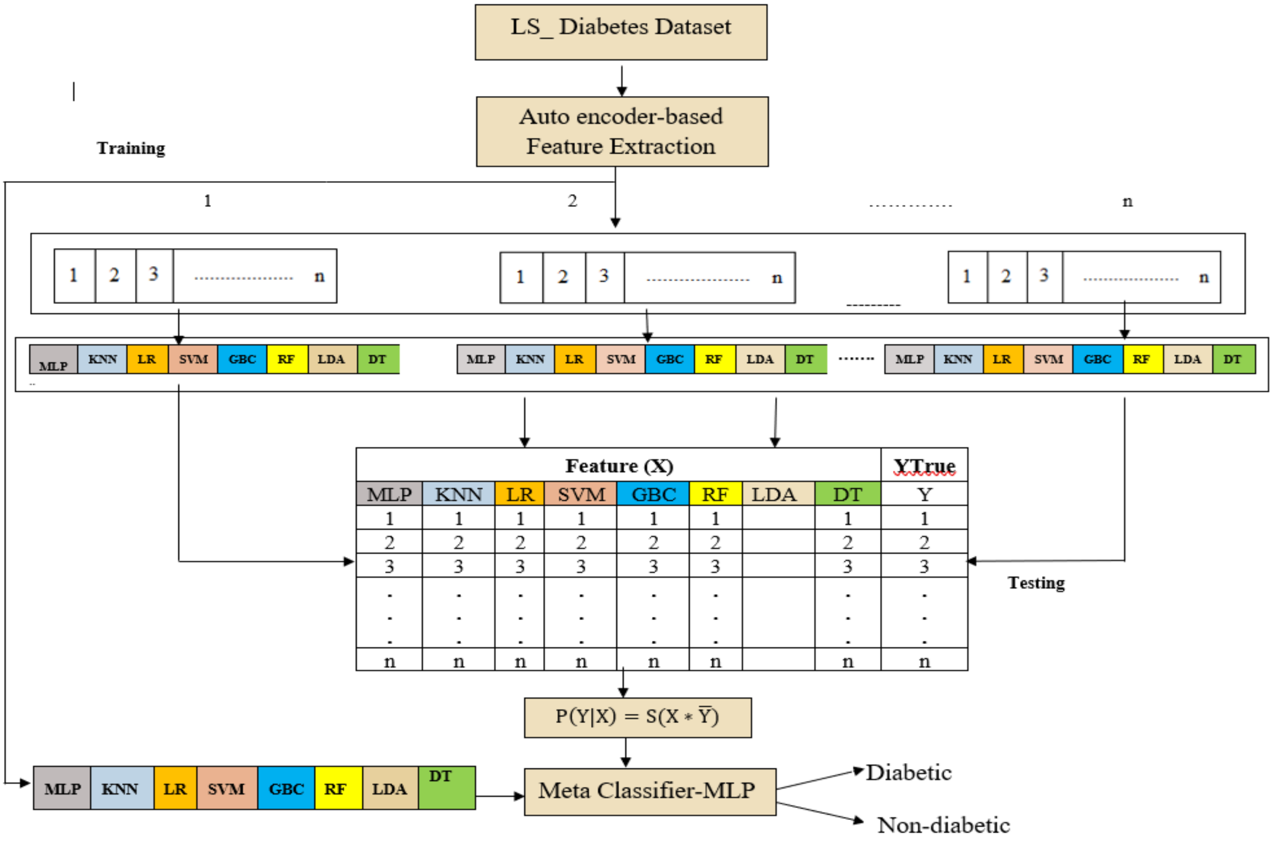

FIGURE 1. Improved ensemble learning with dimensionality reduction (IELDR) model

collected using a questionnaire-based survey and referred as Lifestyle stress diabetes dataset. The questionnaire for research was administered using web based method as well as personally through face-to-face interviews. These features are a combination of demographic, stress related, lifestyle related, genetic and other diagnostic results. As the data were collected through a survey, there were no risks imposed to the participants. The volunteer's consent was obtained before administration of the survey questionnaire. Only study related data were collected without any personal identifiers. There were no other ethical concerns related to the study.

Due to data imbalance, the model performance degrades. The performance of the classifier is ensured by using the data balancing as explained in ALGORITHM 1. The TABLE 1 provides information about the data distribution of all three datasets. The dataset consists of both categorical and numerical values from the questionnaire answers. The collected dataset was above 18 years age group. As most of the features were categorical, hence label encoding was used.

ALGORITHM 1 : Adaptive Synthetic-Tomek Link[25]

```
Input : Imbalanced dataset- Si majority samples, Sj minority samples, Ss synthetic samples, K -k nearest neighbors
Output: BD- Balanced dataset
1: Start
2: Calculate the ratio minority to majority instance ratio
   $\boldsymbol{d = S_j/S_i}$
3:Calculate the synthetic minority samples
   $\boldsymbol{S_s} = (\boldsymbol{S_i - S_j})B$
4: Data generated for each neighbor considering K=3
5: If the sample from minority class is the random nearest neighbor, then create a Tomek Link otherwise remove that sample
6: End
```

The performance of the classifier is ensured by using the data balancing technique -Adaptive Synthetic-Tomek Link [25] as follows:

$$I_i = X_i(X_{si} - X_i) \tag{1}$$

The Eq. (1) represents the synthetic samples generated by selecting λ the random samples. Here X_{si} , X_i are the minority samples. Suppose d(S_i , S_j) is the Euclidian distance of S_i & S_j , here S_i belongs to minority class and S_j samples belongs to majority class. If there are no samples S_k satisfies the condition d(S_i , S_k) < d(S_i , S_j) or d(S_j , S_k) < d(S_i , S_j) then the pair (S_i , S_j) is a Tomek link pair.The algorithm 1 generated by iterating each instance and searching using KNN.

B. PROPOSED MODEL: AUTOENCODER BASED OPTIMAL FEATURE EXTRACTION

The proposed IELDR model worked using an autoencoder-based feature extraction method. Autoencoder is used for a compressed representation of input features. The autoencoder is made up of an encoder and a decoder. The encoder learns to interpret and compress the input to a bottleneck layer-defined internal representation. The decoder takes the encoder's output and attempts to recreate the input sample. Once the autoencoder has been trained, the decoder is discarded, and the encoder is used to compress data and train a model. The IELDR framework refers to stacking responsible for merging the results of base models to achieve more accurate and reliable predictions. The base layer and the meta layer are the two layers of the framework. In the IELDR framework, we trained the base learner in level 0 on the complete dataset and different blocks of the dataset using k-fold. Here, we considered K=10. After training, the probabilities of each class were predicted. The proposed IELDR model trained the base model on different blocks of data, stores a prediction, and then combines the predictions from the base learner into a single vector and passes this vector as input to the meta learner. Meta learner was trained for final prediction using k-fold cross-validation. The existing stacking model trained the models on a complete dataset and combined the predictions using a meta-learner. The experiment used various base learner models at level 0: RF, SVM, LR, DT, GBC, MLP, KNN, and LDA. The multilayer perceptron was used as a meta-classifier at level 1, which showed the final prediction. FIGURE 1 represents the IELDR framework. The IELDR model is divided into two phases: 1) Feature Extraction 2) Learning Ensemble model. At the last model, First, generate predictions from the base models and provide the obtained predictions to the super learner S to create the ensemble prediction on testing or new data samples. Mathematical model for proposed approach is shown in ALGORITHM 2.

ALGORITHM 2. Mathematical model for ILDER- Two levels stacking super learner model with autoencoder for feature extraction

```
Input: Training Dataset: Dtrain = {features(fi), Class(yi)}
Output: Stacking ensemble classifier with super learner S
        Level 0: Feature Extraction using autoencoder

1       Build the autoencoder using the encoder network and the
        decoder model
                $\alpha: \mathcal{D} \rightarrow \mathcal{F},$                                    (1)
                $\beta: \mathcal{F} \rightarrow \mathcal{D},$                                    (2)
            $\alpha, \beta = \underset{\alpha,\beta}{argmin} \parallel \mathcal{D} - (\alpha \circ \beta)\mathcal{D} \,||^2$        (3)
        Encoder model function $\alpha$ in Eq. (1) maps actual data D to
        Latent feature space F
        Decoder model function $\beta$ in Eq. (2) maps the latent feature
        space F to Output D
2       The encoder model Z defined using
                    $z = \sigma(\mathscr{w}\mathfrak{f} + b)$                           (4)
        σ - activation function(sigmoid),
        W - Weight vector
        f -input feature, b -bias
3       The decoder model f' defined
                    $\mathfrak{f}' = \sigma'(\mathscr{w}'z + b')$                         (5)
         F'- decoder model , σ' - activation function(sigmoid)
        W' - Weight vector ,z -latent feature ,b'- bias
4       Calculate the error function E for encoder and decoder network
        using backpropagation method
        $\mathbb{E}(\mathfrak{f}, \mathfrak{f}') = \parallel \mathfrak{f} - \mathfrak{f}' \parallel^2 = \parallel \mathfrak{f} - \sigma'(\mathscr{w}'(\sigma(\mathscr{w}\mathfrak{f} + b)) + b') \parallel^2$   (6)
        Level 1:   Stacking Super Learner Model
5       Select a 10-fold split of the training dataset Dtrain
6       Identify the appropriate list of base Models M
7       Train the first-level M base learners on the training dataset
        for m -> 1 to M  do
            i.  Perform 10-fold cross-validation on each base learner bm on
            Dtrain
            ii. Collect the cross-validated predicted values from each of the
            M base algorithms.
        end for
8       Build the new dataset with the R cross-validated predicted
        score obtained from each M base algorithm and construct a new
        dataset with the shape as    R x M.
        for m->1 to M do
            Build the new dataset as Dnew = ({features(f_newi), Class(yi)}
             where f_newi  = {  b1(fi), b2(fi), b3(fi), ……., bm(fi))
        end for
9       Build the second-level meta-model
          y=g(f) = B(f)= b' (b1(fi), b2(fi), b3(fi), …, m(fi))              (7)
10      Prediction on new data
```

III.RESULTS

This research used different approaches using evaluation metrics like accuracy, recall, precision, f1-score, Mathew correlation coefficient, and roc-auc for the performance evaluation. The proposed IELDR model achieved the highest accuracy of 98.67%. The model showed the lowest error rate of 1.33%. Compared with MLP and SVC, IELDR improved accuracy by 8%, LR by 10.67%, and 12% improvement to LDA shown in FIGURE 2. TABLE 2 illustrates the comparative performance measures of various classifiers, including MLP, SVC, LR, LDA, RF, KNN, GBC,DT and IELDR classifiers.

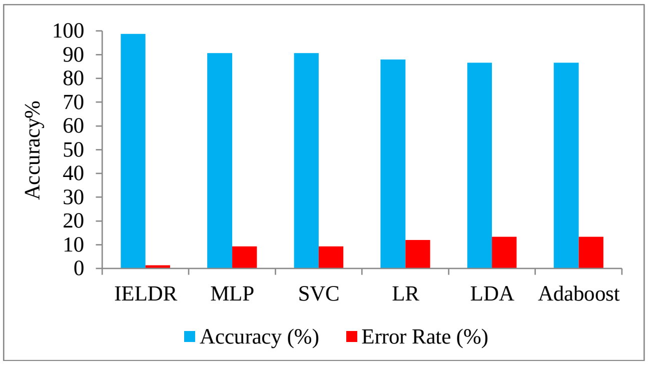

FIGURE 2. Accuracy and error rate comparison of best five model

The IELDR classifier yielded the highest accuracy of 98.67%. If the precision and specificity are considered, the SVC, RF, and KNN classifiers yield a value of 100%. If the recall is considered, the IELDR classifier yielded 100%. The IELDR classifier provided the better result of the f1-score value of 97.56%, while the available algorithms MLP, SVC,

TABLE 2
Comparison of ieldr model with different classifier on various measures

Model	Accuracy	Precision	Sensitivity	Specificity	f1 - score	ROC	Log_Loss	Mathew_ Corrcoef
IELDR	**98.67**	95.24	**100**	98.18	**97.56**	**99.09**	**46**	**96.7**
MLP	90.67	84.21	80	94.55	82.05	87.27	32.2	75.8
SVC	90.67	100	65	100	78.79	82.5	32.2	75.94
LR	88	92.31	60	98.18	72.73	79.09	41.4	67.97
LDA	86.67	81.25	65	94.55	72.22	79.77	46.1	64.28
RF	85.33	100	45	100	62.07	72.5	50.7	61.24
KNN	85.33	100	45	100	62.07	72.5	50.7	61.24
GBC	84	78.57	55	94.55	64.71	74.77	55.3	56.23
DT	74.67	52.17	60	80	55.81	70	87.4	38.36

and LR provided an accuracy of 82.05%, 78.79%, and 72.73%. LDA and Adaboost rendered just 72.22%. The discussed IELDR classifier yielded better results regarding ROC value, Log_loss, and MCC 0.9909, 0.46, and 0.967, respectively.

A. COMPARATIVE ANALYSIS WITH PIMA DATASET AND DIABETES_2019 DATASET

The IELDR model was evaluated on the PIMA dataset from the UCI repository and the Diabetes_2019 dataset collected by the researcher. TABLE 3 shows the comparative results of the IELDR model using the LS_diabetes dataset Vs. PIMA dataset and Diabetes_2019 dataset.

TABLE 3
COMPARATIVE ANALYSIS OF DATASET

Evaluation Metrics	Dataset (Result %)		
	PIMA_Diabetes[23]	Diabetes_2019 [24]	LS_Diabetes
Accuracy	80.71	96.85	98.67
Precision	82.09	96.89	95.24
Sensitivity	59.78	98.42	100
Specificity	92.59	98.42	98.18
F1 score	69.18	97.65	97.56
ROC	76.19	96.08	99.09
MCC	57	92.9	97

Best results were observed on the LS_diabetes dataset with accuracy, precision, sensitivity, specificity, f1-score, roc, and MCC value by 17.96%, 13.15%, 40.22%, 5.59%, 28.38%, 22.09%, and 0.4% compared to the PIMA dataset. The IELDR model achieved a 98.67% accuracy on the LS_diabetes dataset,while Diabetes_2019 showed an accuracy of 96.85%. The IELDR model performed the best result on the LS_diabetes dataset with performance measures such as accuracy, sensitivity, roc, and MCC value by 1.82%, 1.58%, 3.01%and 0.04 % compared to the Diabetes_2019 dataset.

CASE I: Comparison IELDR model and the best model experimented on the PIMA.

A comparison of the IELDR model and the best model experimented on the PIMA Indian diabetes dataset from state of the art is shown in TABLE 4.

TABLE 4
COMPARATIVE RESULTS OF IELDR MODEL AND OTHER MODELS ON PIMA DATASET (Case1- IELDR Vs. PIMA [23])

Evaluation Metrics	IELDR	[26]	[27]	[34]
Accuracy	80.71	79.08	75	76.3
Precision	82.09	73.13	47	75.9
Specificity	92.59	83.44	76	NA

The result showed that accuracy, precision and specificity rise by 1.63%, 6.19% and 9.15% from the existing researcher. Observed result showed that the proposed IELDR achieved the highest accuracy, precision and specificity values of 80.71%, 82.09% and 92.59%, respectively.

TABLE 5
IELDR results versus methodology used on diabetes_2019 dataset

Case2- Diabetes_2019 [24]			
Evaluation Metric	**IELDR**	**[24]**	**[18]**
Accuracy	**96.85**	94.1	96.81
Precision	96.89	97.6	96
Sensitivity	**98.42**	94.3	92.3
Specificity	98.42	93.4	98.52
F1- score	**97.65**	95.9	94.11

CASE II: Proposed IELDR based framework and the framework used by existing work on the diabetes_2019 dataset.

The proposed IELDR-based framework and the framework used by existing researchers on the diabetes_2019 dataset were compared. The IELDR framework showed the excellent result shown in TABLE 5. The IELDR model yielded the highest accuracy, precision, sensitivity, and f1-

score with a value of 96.85%, 96.89%, 98.42%, and 97.65%, respectively. Using the IELDR model, we achieved an accuracy of 80.71%, 96.85%, and 98.67% on the PIMA dataset, Diabetes_2019 dataset, and LS_diabetes dataset, respectively.

IV.DISCUSSIONS

In this study, we used the IELDR model to predict the development of type 2 diabetes based on lifestyle and stress. In this research, the IELDR classifier provided the highest accuracy. This accuracy was higher than other classifiers, including MLP, SVC, LR, LDA. These results could be ascribed to an autoencoder-based feature extraction method in contrast to the conventional feature selection method. We also used two level stacking model with MLP as a base learner. The study showed the highest sensitivity with the IELDR model. The IELDR classifier also showed better ROC, Log_loss, and MCC values than other models. In the next step, we used the IELDR model to LS_Diabetes dataset and PIMA dataset to compare the results. We observed better results of the IELDR model on the LS_Diabetes dataset than the PIMA dataset. Interestingly, our dataset was smaller than the PIMA dataset [23], which points to the robustness of the IELDR model we used. We achieved better accuracy, precision, sensitivity, specificity, F1 score, ROC, and MCC than the PIMA dataset. Several other authors have also tested their methodology on the PIMA dataset.

Ahmed et al. [18] used seven classifiers to create a system for predicting diabetes. These authors used two datasets (PIMA and Tigga and Garg). The PIMA dataset achieved the highest accuracy for support vector machine (SVM) and random forest (RF) with 80.26 %; the other dataset achieved the highest accuracy with a decision tree (DT) and RF with 96.81 % and developed a web app. Sivashankari et al. [19] designed a stacked ensemble model on PIMA dataset showed accuracy, precision, recall, and F1-score values of 93.1%, 84%, 83.9%, and 83.5%, respectively. Diwani and Sam [20] developed a system on the PIMA dataset showed Naïve Bayes(NB) performed better with an accuracy of 76.30%. Krishnamoorthi et al. [21] Using ML, the author proposed an intelligent diabetes mellitus prediction framework (IDMPF). Mahabub et al. [22] designed a system to predict diabetes by improving accuracy using 11 classifiers on the PIMA dataset. An ensemble voting classifier was developed using SVM, MLP, and KNN by applying hyper parameter tuning and cross-validation provided an accuracy of almost 86%.

A diabetes prediction model was created by Tigga and Garg [24]. A dataset comprising 952 instances and 18 attributes related to lifestyle, health, and family background. Comparative analysis done with PIMA dataset. RF, showed an accuracy of 94.10% for the collected data and 75% for the PIMA dataset. Kumari et al. [26] designed an ensemble soft voting approach. PIMA dataset and breast cancer datasets were used for evaluation purposes. In this study, the authors used an ensemble of three algorithms: RF, LR, and Naïve Byes. They compared performance with an existing system and found that the approach outperforms the with 79.08% accuracy and an F1-score value of 80.6% on the PIMA dataset. Our results were comparatively better than those reported by Kumari et al. [26]

Chatrati S.P. et al. [27] have reported findings of an application for predicting diabetes and hypertension. The authors said the SVM classification algorithm is the most accurate. Patil and Shah [28] developed a stress-based model using stress-based and demographic features. The stress prediction model used RF, LR, and SVM classifiers. In this study, SVM provided the best accuracy 80.17% compared to other classifiers. Saxena et al. [29] performed a comparative study using four classifiers, MLP, DT, RF, and KNN, with Correlation, Information Gain, and PCA feature selection techniques on the PIMA dataset. Hyperparameter optimization and pre-processing methods were used. They contrasted the results with and without feature selection and discovered an accuracy rate of 79.8% with RF. Kannadasan et al. [30] proposed a Deep Neural Network (DNN) based framework using stacked autoencoders. The proposed framework was experimented on the PIMA dataset. With an accuracy of 86.26 %, this model outperformed other models. Kiranashree B.K. et al. [31] introduced a system for stress detection based on physiological factors using ML. NB, RF, and SVM models showed the best accuracy. SVM provided an accuracy of 96.67% compared with other classifiers. Ahuja and Banga [32] performed a student-centric study to evaluate mental stress before examining and spending time online. The dataset used for this study was 206 records from JIIT. Four classifiers used for the study were linear regression, NB, RF, and SVM. SVM provided a specificity, accuracy, and sensitivity value of 100%, 85.71%, and 75%, respectively Ayush and Divya [33] designed a predictive model based on personal indicators for establishing a relationship between lifestyle activities and the risk of diabetes. Lifestyle activities included sleeping, eating, physical activities, BMI, and waist circumference. The collected dataset contained a total of 180 records. The CART model showed an accuracy rate of 75%. Sisodia D. et al. [34] designed a system to predict the possibility of diabetes by achieving higher accuracy. PIDD dataset. The performance results for NB were the best, with a maximum accuracy of 76.3% and the highest ROC value of 81.9%. We observed better results than these studies [26-27,34].

In addition, we also used our model on the Diabetes_19 [28] dataset obtained from the researcher with the request. Our algorithm faired on the accuracy, sensitivity, roc, and mcc value. We also validated our methodology with the other researcher's methodology [18, 28]. Our results of accuracy, precision, sensitivity, and f1-score were better. Specificity value was also approximately similar to these researchers [18, 28]. The results of our study could be useful to the researchers for early prediction of type 2 diabetes. Early prediction may help to manage the disease more effectively. The insights of the study results can help general population as well as researchers to understand the risk factors for development of type 2 diabetes and work on them to avoid the disease related complications. Our study has some limitations. Smaller database with convenience sampling method is one of the limitations of our study. Cross-sectional method of data collection is another limitation of our study. Use of more algorithms for comparison may further add more insights about the results.

V.CONCLUSION

The study aimed to build a machine learning based risk prediction model. The proposed IELDR approach was designed using a two-level stacking autoencoder-based model showed 98.67% accuracy. The proposed framework was validated with two datasets, the PIMA Indian diabetes dataset and the Diabetes_2019 dataset, for its stability showed an accuracy of 80.71% and 96.85% respectively. The IELDR model achieved better accuracy, precision, sensitivity, specificity, f1 score, roc, and MCC than the PIMA dataset. Similarly, compared to the Diabetes_2019 dataset, the IELDR model provided better accuracy, sensitivity, roc, and MCC values on the LS_Diabetes dataset. Implementation and assessment of the proposed IELDR showed its usefulness in predicting type 2 diabetes. This model can help to predict the possibility of developing type 2 diabetes by utilizing a comprehensive India-based LS_diabetes dataset. The IELDR model can be a promising tool for analysing and forecasting type 2 diabetes using autoencoder-based feature extraction with a two-level stacking model.

Considering the rising recognition of ML in healthcare, the findings of this study may help predict type 2 diabetes. Overall, the results of our research and validation with other datasets provide significant insights for predicting type 2 diabetes development. Stress and lifestyle should be considered vital risk factors for the development of DM, and appropriate strategies should be designed to control these risk factors. Further studies can be conducted on large dataset with randomized sample selection from different states of the country and results can be compared with proposed datasets.

CONFLICTS OF INTEREST

The authors reports that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

The paper background work, methodology, dataset collection, implementation, result analysis and comparison done by first author, conceptualization done by second author, preparing and editing draft, visualization have been done by first and third author. The supervision, review of work has been done by fourth and fifth author.

ACKNOWLEDGMENTS

Authors wish to thank Dr Deepak Bhosle for their guidance for understanding of medical aspects of the disease and overall process of writing the manuscript.

REFERENCES

[1] Anant, Mohamad Goldust, and Uwe Wollina,"Herpes zoster: A Review of Clinical Manifestations and Management" *Viruses* Vol.14, No.2,pp.192,2022 .

[2] Patil S, Patil A," Systemic lupus erythematosus after COVID-19 vaccination: A case report", *J Cosmet Dermatol,*Vol.20, No.10,pp.3103-3104,2021.

[3] Pavate, A., Bansode, R. , "Design and Analysis of Adversarial Samples in Safety–Critical Environment: Disease Prediction System",In: *Lecture Notes in Computational Vision and Biomechanics*, vol 37. Springer, Singapore.

[4] A. Pavate and N. Ansari, "Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques," In: *Fifth International Conference on Advances in Computing and Communications (ICACC)*, Kochi, India, 2015, pp. 371-375,2015.

[5] Pavate, A., Nerurkar, P., Ansari, N., Bansode, R.,"Early Prediction of Five Major Complications Ascends in Diabetes Mellitus Using Fuzzy Logic. In: *Soft Computing in Data Analytics. Advances in Intelligent Systems and Computing*, vol 758. Springer, Singapore.

[6] Aruna Pavate., et al. "Diabetic Retinopathy Detection-MobileNet Binary Classifier ," *Acta Scientific Medical Sciences* Vol.4.No.12 pp.86-91,2020.

[7] Patil, Rohini, Kamal Shah, and Deepak Bhosle. "Impact of COVID-19-related Stress on Glycaemic Control in Hospitalized Patients with Type 2 Diabetes Mellitus." *Journal of Health Sciences & Surveillance System* Vol.10, No. 4 pp. 397-402,2022.

[8] Patil, R., Shah, K. ,"Performance Evaluation of Machine Learning Classifiers for Prediction of Type 2 Diabetes Using Stress-Related Parameters". In: *Data Science and Security. Lecture Notes in Networks and Systems*, vol 462. Springer, Singapore,2022.

[9] Tsang KCH, Pinnock H, Wilson AM, Shah SA. ,"Application of Machine Learning Algorithms for Asthma Management with mHealth: A Clinical Review", *J Asthma Allergy*. Vol.29, No.15, pp. 855-873,2022.

[10] Dritsas E, Trigka M,"Stroke Risk Prediction with Machine Learning Techniques", *Sensors,* Vol.22,No.13, pp.4670,2022.

[11] Kim, J., Mun, S., Lee, S. et al.," Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea", *BMC Public Health* Vol.22, No.664 ,2022.

[12] R. Indrakumari, T. Poongodi, Soumya Ranjan Jena, "Heart Disease Prediction using Exploratory Data Analysis"In:*Procedia Computer Science,* Vol.173, pp.130-139 2020.

[13] Diabetes http://www.who.int/en/news-room/fact-sheets/detail/diabetes accessed on 21st Feb 2019

[14] IDF SEA members https://www.idf.org/our-network/regions-members/south-east-asia/.../94-india.html assessed on 22nd Feb 2019

[15] ICMR Guidelines for management of type 2 diabetes 2018 https://main.icmr.nic.in/sites/default/files/guidelines/ICMR_GuidelinesType2diabetes2018_0.pdf accessed on 23 July 2021

[16] Patil R, Shah K.,"Machine learning in healthcare: Applications, current status, and future prospects",In: *Handbook of Research on Machine Learning: Foundations and Applications,* (1st ed.). Apple Academic Press. 4 August 2022.

[17] Svalastog AL, Donev D, Jahren KN, et al. ,"Concepts and definitions of health and health-related values in the knowledge landscapes of the digital society", *Croat Med J.* Vol.58,pp.431-435,2017.

[18] Ahmed N, Ahammed R, Islam M, et al. ,"Machine learning-based diabetes prediction and development of smart web application",In: *International Journal of Cognitive Computing in Engineering,* Vol. 2,pp.229-41,2021.

[19] Sivashankari R, Sudha M, Hasan MK, et al.," An empirical model to predict the diabetic positive using a stacked ensemble approach", *Front. Public Health*; Vol.9, 2022.

[20] Diwani SA. Sam A.," Diabetes forecasting using supervised learning techniques",*Adv. Comp. Sci. Int. J*,Vol.3,pp.10-18,2014.

[21] Krishnamoorthi R, Joshi S, Hatim Z, et al.,"A novel diabetes healthcare disease prediction framework using machine learning techniques", *Journal of Healthcare Engineering*, 2022.

[22] Mahabub A.,"A robust voting approach for diabetes prediction using traditional machine learning techniques", *SN Applied Sciences*, Vol. 1No.1667,2019.

[23] Pima Indians Diabetes dataset. Available from: http://archive.ics.uci.edu/ml/machine learning-databases/pima-indians-diabetes/pima-indians-diabetes data. Accessed: 1st May 2008.

[24] Tigga N, Garg S.,"Prediction of type 2 diabetes using machine learning classification methods classification",In:*International*

*Conference on Computational Intelligence and Data Science, Procedia Computer Science,*Vol.167,pp.706-16,2020.

[25] Ullah, N. Javaid, M. U. Javed, Pamir, B. S. Kim, and S. A. Bahaj, "Adaptive Data Balancing Method Using Stacking Ensemble Model

[26] and Its Application to Non-Technical Loss Detection in Smart Grids," *IEEE Access*, vol. 10,pp. 133244–133255, 2022.

[27] Kumari S, Kumar D, Mittal M.," An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier", *International Journal of Cognitive Computing in Engineering*, Vol. 2,pp.40-46, 2021.

[28] Chatrati SP, Hossain G, Goyal A, et al.," Smart home health monitoring system for predicting type 2 diabetes and hypertension", *J. King Saud Univ. Comput. Inf. Sci,* 2020.

[29] Patil R, Shah K.," Assessment of risk of type 2 diabetes mellitus with stress as a risk factor using classification algorithms",In: *International Journal of Recent Technology and Engineering,* Vol. 8,pp. 11273–77,2019.

[30] Saxena R, Sharma SK, Gupta M, et al.," A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods", *Computational Intelligence and Neuroscience,* 2022.

[31] Kannadasan K, Edla D, Kuppili V.,"Type 2 diabetes data classification using stacked autoencoders in deep neural networks", *Clinical Epidemiology and Global Health*, pp.530-35, 2019.

[32] Kiranashree BK , Ambika C, Radhika AD. ,"Analysis on machine learning techniques for Stress detection among employees",*Asian Journal of Computer Science and Technology,*Vol.10,pp. 35-7, 2021.

[33] Ahuja R. Banga A.," Mental stress detection in university students using machine learning algorithms",In:*International Conference on Pervasive Computing Advances and Applications (PerCAA 2019), Procedia Computer Science,* Vol.152,pp. 349-53, 2019.

[34] Anand A, Shakti S.," Prediction of diabetes based on personal lifestyle indicators",In :*1st international conference on next generation computing technology:IEEE,*2015.

[35] Sisodia D. Sisodia D.," Prediction of diabetes using classification algorithms" In: *Procedia Computer Science*, Vol. 132,pp.1578–1585, 2018.

DR ROHINI PATIL is currently working as Assistant Professor in Terna Engineering College, Navi Mumbai. She has completed her Master's in Computer Engineering and PhD in Information Technology from Mumbai University. She has published more than 25 papers in international journals/conferences. Her area of interest includes Machine Learning, Data Mining and Data Science. She has published three Patents.

DR ANANT PATIL is working as Associate Professor in Dr DY Patil Medical College, Navi Mumbai. He is MBBS and MD Pharmacology by qualification. He has published more than 140 articles in national and international journals related to diversified disease areas.

PROF SUREKHA JANRAO is working as an Assistant Professor in K.J.Somaiya Institute of Technology; Mumbai. She has done her Master's and PhD. in computer engineering from Mumbai University. She has published more than 10 papers in international journals. Her area of interest includes Machine Learning, Data Mining and IoT. She has published two international and Indian Patent.

DR SANDIP BANKAR received M.E. degree in Computer Engineering and Ph.D. degree in Information Technology from University of Mumbai in India. He is currently working as an Assistant Professor in computer engineering at NMIMS deemed to be university, Mumbai. His current research interest includes Machine Learning and Blockchain.

DR KAMAL SHAHhas a vast teaching experience and more than 10 years of research experience. She has completed her Ph.D. in engineering in 2010 in the field of image processing. Currently she is working as In-Charge Principal, St. John College of Engineering and Management, Palghar, Mumbai. She has more than 40 research papers to her credit and she has completed post-doctoral research in the field of Quantum computing from BARC, Mumbai. Her area of interest includes Block chain technology.