I.J. Intelligent Systems and Applications, **2022, 4, 11-20**
Published Online August 2022 in MECS (http://www.mecs-press.org/)
DOI:10.5815/ijisa.2022.04.02

A Novel Hybrid Approach for Detection of Type-2 Diabetes in Women Using Lasso Regression and Artificial Neural Network

Yogendra Singh
Department of Electronics and Communications, University of Allahabad, Prayagraj, India
E-mail: sachin12345jan@gmail.com

Mahendra Tiwari
Department of Electronics and Communications, University of Allahabad, Prayagraj, India
E-mail: mahendra@allduniv.ac.in

Received: 02 January 2022; Accepted: 23 March 2022; Published: 08 August 2022

Abstract: Diabetes is a life-threatening and long-lasting illness that produces high blood glucose levels. Diabetes may cause various diseases, including liver disease, blindness, amputation, urinary organ infections, etc. This research work aims to introduce a hybrid framework to enhance outcomes predictability and interoperability with reduced ill-posed problems, over-fitting problems, and class imbalance problems for diagnosing diabetes mellitus using data mining techniques. Diabetes may be recognized in many ways. One of these methods is data mining techniques. The use of data mining to medical data has yielded meaningful, significant, and effective results that may improve medical expertise and decision-making. This study suggests a hybrid technique for detecting DM that combines the lasso regression algorithm with the artificial neural network (ANN) classifier algorithm. The Lasso regression technique is used for variable selection and regularization. Because the dataset was shrunk, the computing time was considerably minimized. The ANN classifier received the Lasso regression output as an input and classified patients correctly as diabetic and non-diabetic, i.e., tested positives and negatives. The Pima Indians dataset was used in this experiment, consisting of 768 samples of female participants who are diabetic and non-diabetic. According to experimental observations, the proposed hybrid technique achieved 93% classification accuracy for predicting diabetes mellitus. The experimental results showed that our proposed method had a classification accuracy of 93% for determining whether a patient has diabetes or not. The experimental outcomes demonstrated that a hybrid data-mining approach might assist clinicians in making better diagnoses when identifying diabetes patients.

Index Terms: Diabetes mellitus, Lasso regression, artificial neural network, Deep learning, Predictive model.

1. Introduction

Diabetes affects around 422 million people globally, according to the World Health Organization [1]. According to a study released by the International Diabetes Federation (IDF) in 2019, China has the highest number of diabetes patients globally, with diabetes patients aged 20 to 79 years, and is foreseen to continue to happen in 2045 [2]. According to the research, Mauritius is also ranked first in Southeast Asia for age-adjusted comparative diabetes prevalence among persons aged 20 to 79 years old. In India, Bangladesh, and Sri Lanka, diabetes affects 98.9% of the adult population. Particularly in India, where diabetes affects 77 million individuals.

Diabetes mellitus is a chronic sickness or metabolic disorder in which the glucose levels in the human blood surpass normal levels [3]. High blood glucose, often known as hyperglycemia, is caused by inadequate insulin production and activity in the human body. Diabetes causes a variety of diseases, including retinopathy, heart attack and stroke, nephropathy, neuropathy, skin diseases (such as bacterial and fungal infections), hearing problems [4,5], vision loss, kidney failure or renal disease, congestive heart failure, stroke, and lower limb amputation. According to the WHO, over 4.2 million people worldwide are expected to die from diabetes and its complications in 2019, with diabetes accounting for 2% of all fatalities in India. Consequently, early diabetes detection and prevention are crucial for saving human lives.

Machine learning is a beacon of hope in the direction of the treatment of diabetes. It assists medical specialists in identifying, diagnosing, predicting, and classifying patients and achieving the accuracy, adroitness, and reliability of the healthcare system. A lot of research is being done to overcome these problems and quickly and accurately diagnose and analyze diabetes. But most of the conversational models suffer from performance-related issues like bias-result, over-fitting, under-fitting, class imbalance problems, attribute selection, time-consuming, etc. Consequently, we proposed a new hybrid model to enhance outcomes predictability and interoperability with reduced ill-posed problems, over-fitting problems, and class imbalance problems and compare its results with four machine learning algorithms named Naive Bayes, K-Nearest Neighbor, Support Vector Machine, and the J48 algorithm. This study is primarily concerned with the performance, theoretical, characteristics, and algorithm approaches. For disease prediction, the classification technique was used rather than regression. A hybrid approach combining lasso regression and an artificial neural network has been developed to identify diabetic patients. When compared to base classifiers, the hybrid technique has been shown to be more effective. The performance of the method, as mentioned above, was tested using the following criteria: accuracy, precision, recall, F1-score, and ROC.

The following parts comprise this research paper: Section 2 covers relevant work in this field. Section 3 explains the proposed technique. Section 4 discusses the findings and analysis of the proposed methodology, and the aforementioned methodology's results have been compared and analyzed with traditional ML algorithms and state-of-the-art techniques. Finally, Section 5 summarizes the findings of our research work.

2. Related Works

Various researchers have conducted several studies using different data mining techniques and machine learning algorithms on various healthcare datasets to develop a prediction model for the health care sector, where analysis and predictions are carried out using different methods and techniques.

J.J. Khanam and S.Y. Foo used the PID dataset to develop diabetes prediction models with three approaches (Neural Network, Machine Learning, and Data Mining). In this research study, the author resolved missing values in the dataset with suitable mean values, detected various outliers and removed extreme values based on interquartile ranges, used the Pearson’s correlation method to choose the most effective attributes, and applied normalization to transform the dataset into an appropriate form. They observed that the prediction model based on Support Vector Machine (SVM) and Logistic Regression (LR) performed extremely well in their suggested model. The NN model was applied to build a forecasting model with three distinct hidden layers and varied epochs, and it was observed that two hidden layers and a 400-epochs-based NN model gave an overall accuracy of 88.6% [6].

S. Kumari et al. proposed a model that they were able to create a diabetes prediction model with better accuracy by utilising the Pima India Diabetes dataset by implementing a combination of machine learning algorithms. They utilized a combination of three supervised algorithms, namely Naive Bayes (NB), logistic regression (LR) and random forest (RF), to create an ensemble soft voting classifier to determine diabetes or non-diabetes. On the PID dataset, their proposed model performs exceptionally well in terms of performance measures, with accuracy of 79.04%, precision of 73.48%, recall of 71.45%, and F1 score of 80.6%. The expected methodology's efficacy has also been distinguished and investigated using a breast cancer dataset. On the breast cancer dataset, the suggested model (ensemble soft voting classifier) obtained 97.02 overall accuracy [7].

In 2017, Tao Zheng et al., performed a study and introduced a data-informed system to recognize patients of Type 2 Diabetes from Electronic Health Records with the help of feature engineering and machine learning. They also showed the identification performance of other machine learning algorithms. The main intention of this research is to minimise the impact of missing data in order to identify more Type 2 diabetes patients. The highest performing algorithm on the engineered features is employed and the results reveal that the framework has an AUC of around 0.98 [8].

In 2020, H. Naz and S. Ahuja claimed in their published paper that deep learning is efficient and convenient for developing predictive models in the healthcare sector. The study recommended a new approach for diabetes prediction depending on a variety of machine learning techniques using the PIMA dataset. The author used four distinct classifier techniques, including Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT), and Deep Learning (DL), to get promising findings with accuracy ranging from 90% to 98%. Deep Learning, when compared to ANN, NB, and DT, produces the greatest results on the PID dataset, and the accuracy was 98.07% [9].

T.M. Le et al. conducted a study in 2020 to design a high-efficiency predictive model based on machine learning techniques to recognise patients with diabetes mellitus. In order to construct a predictive model, the authors have utilized wrapper-based feature selection utilizing Grey Wolf Optimization (GWO) and Adaptive Particle Swam Optimization (APSO) to optimize the multilayer perception and reduce the number of required input attributes. The following six classifier algorithms: Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN), Navies Bayes (NB), Random Forest (RF), and Logistic Regression (LR) based models are compared with the proposed diabetes prediction model. The investigation and comparison were performed on several evaluation matrices, and they revealed that the proposed predictive model not only used fewer attributes but also helped to achieve higher accuracy (96% for GWO-MLP and 97% for APGWO-MLP) in respect to other classification techniques [10].

Sneha and T. Gangil carried out research in 2019 to create a predictive machine learning algorithm based on significant characteristics and to choose the best classifier to produce the most accurate results in comparison to clinical outcomes. SVM, Naive Bayes, k-NN, decision tree method, and random forest are utilised in the proposed model, with Nave Bayesian result states having the greatest accuracy of 82.30% [11].

3. Materials and Methods

This section describes the workings of the proposed diabetes prediction model. (1) The first phase begins with the collection of data; (2) the second phase is pre-processing, which includes missing value identification, outliers' identification and replacement, random sampling, lasso regression, and normalization. (3) In the third stage, artificial neural networks are used as a classification technique. The suggested model is depicted in Figure 1, and these phases of the prediction framework are described in detail below.

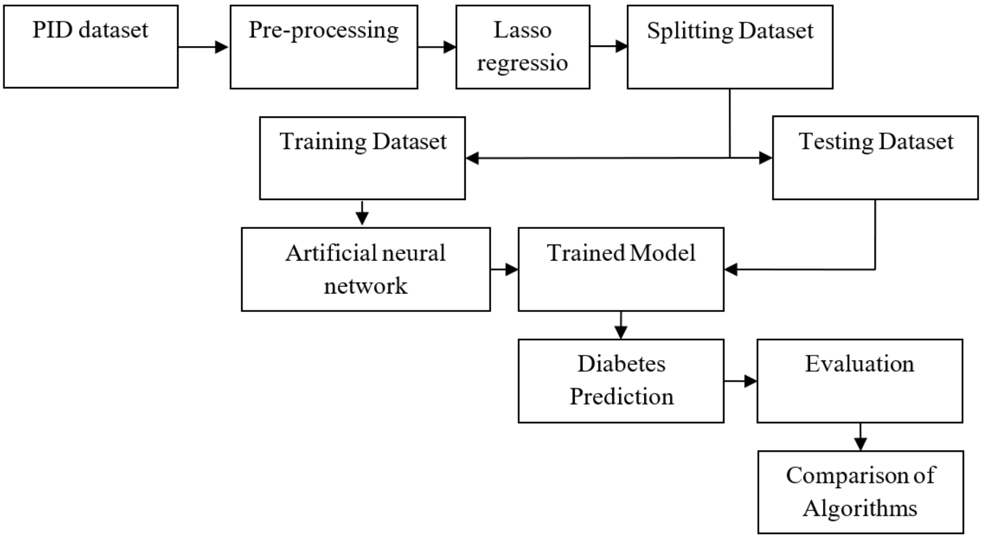

Fig.1. Diabetes prediction System.

3.1. Dataset

The Pima Indian diabetes dataset that has been considered for this experiment is taken from the National Institute of Diabetes and Digestive and Kidney Diseases (publicly available at the kaggle and UCI ML Repository) [12]. The goal of utilising this dataset was to predict whether a patient has diabetes or not, based on the dataset's particular attributes. The Pima diabetes dataset has 8 attributes and one class attribute, with 768 records describing female patients aged between 21 and 81 years [13]. The dataset is described briefly in Table 1.

Table 1. A brief description of the Pima Indians diabetes dataset.

Sr. No.	Feature name	Description
1	Pregnancy	The number of pregnancies a woman has.
2	Glucose	Sugar level presents in blood
3	Blood pressure	Diastolic blood pressure reading (mm Hg)
4	Skin thickness	The thickness of the triceps' skin folds (mm)
5	Insulin	2 h serum insulin
6	BMI	Body Mass Index (weight / height^2 or Kg/m^2)
7	Age	Patient's age
8	Diabetes pedigree function	The possibility of affecting by disease depending on the patient's ancestors' history.

3.2. Pre-processing

Data Pre-processing is one of the major processes in which the data gets transformed in order to build a better machine learning model with high accuracy [6]. Most healthcare-related data has missing values and other impurities that might reduce the effectiveness of the dataset [14]. To improve the quality of data, Data Pre-processing has been conducted, which includes missing value imputation, outliers' identification and imputation, random sampling, normalization, and lasso regression. The description of these techniques is discussed in the below subsections.

A. Missing value imputation

Real-world data frequently includes a lot of missing values. The reason for missing values may be data corruption or failure to record data. The handling of missing data is extremely essential during the preparation of the dataset since many ML algorithms do not accept missing values. In our experimental dataset, missing values are represented by zero [15]. Using Jupyter Notebook, we retrieved the number of missing data points in glucose, blood pressure, skin thickness, insulin, BMI, and age by replacing zero with NaN. We can readily observe the number of missing values found in the dataset in Table 2. Since the number of missing values was very large, we could not remove all records containing missing values as this would have reduced the size of the dataset. Therefore, the missing value was replaced with the appropriate median value by grouping the diabetic and non-diabetic.

Table 2. The number of missing values in PIMA dataset.

Sr. No.	Feature name	No. of missing values	Percentage of missing value
1	Pregnancy	0	0.00%
2	Glucose	5	0.65%
3	Blood pressure	35	4.56%
4	Skin thickness	227	29.56%
5	Insulin	374	48.70%
6	BMI	11	1.43%
7	Age	0	0.00%
8	Diabetes pedigree function	0	0.00%

B. Outlier identification and replacing

Data points that are substantially different from the rest of the observations are called outliers. In other words, they’re uncommon numbers in a dataset. Outliers are a concern in many statistical studies because they may lead tests to irrelevant result or biased actual results. We analysed the dataset using Jupyter Notebook to look for outliers based on the interquartile range. Table 3 shows the number of outliers, which shows that there are a total of 293 outlier values. Unfortunately, there are no definitive statistical criteria for detecting outliers. As a result, we replaced all those outlier values in the PID dataset that were above the 95th and below the 5th percentiles with their respective percentile values. Equations (1) and (2) were used to find outlier values, whereas equations (3) and (4) were used to find extreme values.

$$Q1 - EVF * IQR \leq x < Q1 - OF * IQR \tag{1}$$

$$Q3 + OF * IQR < x \leq Q3 + EVF * IQR \tag{2}$$

$$x < Q1 - EVF * IQR \tag{3}$$

$$x > Q3 + EVF * IQR \tag{4}$$

Where,

Q1 = 5% quartile, Q3 = 95% quartile, IQR = Interquartile Range, difference between Q1 and Q3
OF = Outlier Factor, EVF = Extreme Value Factor

Table 3. Outliers in PID dataset.

Sr. No.	Features Name	Number of Outliers	Percentage of outliers
1.	Pregnancies	3	0.39%
2.	Glucose	0	0.00%
3.	BloodPressure	32	4.17%
4.	SkinThickness	115	14.97%
5.	Insulin	70	9.11%
6.	BMI	20	2.60%
7.	DiabetesPedigreeFunction	38	4.95%
8.	Age	15	1.95%

C. Random Sampling

The total number of entries in the PIMA diabetes dataset is 768, with 500 patients being non-diabetics and 268 being diabetic. The bias result will be produced if a predictive system is trained with an imbalanced class attribute. Random sampling is employed to solve this issue. Random sampling is classified into two types: random under sampling and random over sampling. We utilized random over sampling in our proposed model. The main objective of using this method is to have small dataset size and no information loss.

D. Normalization

Normalization is a feature scaling approach that involves changing the values of numeric columns in a dataset to a similar scale without sacrificing information or compromising variance in value ranges [16]. Differences in scales among input variables may exacerbate the classification problem's complexity. The Z-Score Normalization approach was used to normalize the diabetes dataset. The mathematical formula for normalisation is shown in Eq. (5), where Z is the normalised attribute value, x_i is the initial value of attribute, μ is the mean, and σ is the standard deviation.

$$Z = \frac{x_i - \mu}{\sigma} \tag{5}$$

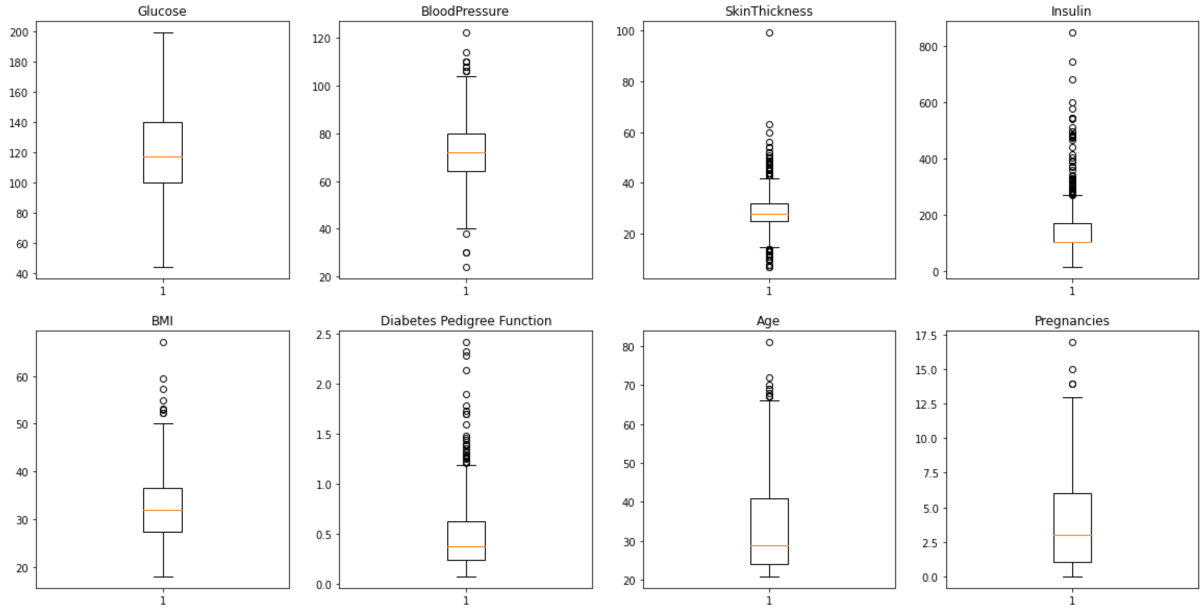

Fig.2. Identified Outliers in PID dataset.

3.3. *Lasso regression*

Lasso is a linear regression method used in feature selection and regularization to improve predictability and interpretability as well as reduce ill-posed problems or over-fitting problems in the machine model. Robert Tibshirani, a professor of statistics, initially proposed the approach in 1996. The LASSO is a subset of the penalized least squares regression using the L1-penalty function.

The Lasso estimate can be defined by

$$\hat{\beta} = arg_{\beta} \min \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\} \tag{6}$$

Which can also be written as

$$\hat{\beta} = arg_{\beta} \min \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{, subject to } \sum_{j=1}^{p} \left| \beta_j \right| \leq t$$

LASSO tries to reduce the size of certain coefficients of model while setting others at zero. Hence, it is a forward-looking variable selection method for regression. It decreases the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. LASSO was initially developed in terms of least squares, but it may be applied to a broad range of models.

By combining the merits of ridge regression with subset selection, LASSO enhances both prediction accuracy and model interpretability. If the set of variables has a strong correlation, LASSO selects just one and reduces the others to zero. It decreases the variability of the estimates by decreasing some of the coefficients precisely to zero, providing easily interpretable models. Because of all these features, we used Lasso to optimize our predictive model.

3.4. *Artificial neural networks*

Artificial neural network (ANN) is a branch of artificial intelligence and a vital technology in data mining which is inspired by the functionality of biological neural networks. Artificial neural network, like the human brain, is interconnected to one another neurons in various layers of networks. It has three types of layers: input, hidden, and output layers. Each node in the present layer gets input from other nodes in the previous layers, and the weights between them are adjusted so that the whole network learns to compute the output [17]. The artificial neural network (ANN) recognises complicated patterns and learns from them. There are billions of neurons in the human brain. Axons link these cells to other cells, and a single neuron is known as a perceptron. Dendrites receive input and interpret it as stimulus. Similarly, the ANN is made up of a number of nodes that are linked together. A weight represents the link between units. The goal of ANN is to transform data into meaningful output [18].

Optimization and parameter selection are crucial in an artificial neural network classifier. As a consequence, the hyperparameters employed in this classifier implementation are described in detail here. The number of hidden layers used in implementation of ANN is four with 1000 epoch and small dropout of 20%. In ANN, the weighted sum of input is processed by the activation function in the hidden layer. In our experiment, we employed the sigmoid and RELU activation functions. The activation function for the input and hidden layers was the Rectified Linear Unit (ReLU), while the sigmoid was the output activation function. The optimizer is required to reduce the output error during the back propagation method. We used Adam as an optimizer. In our experiment, the output was in binary form, so we used binary_crossentropy as a loss function.

3.5. *Proposed algorithm for predicting diabetes*

We presented a hybrid technique combining lasso regression and artificial neural networks. The suggested method first optimizes the dataset before categorizing the patients as diabetes or non-diabetic. Table 4 shows the procedures for implementing the suggested method.

Table 4. Proposed algorithm for diagnosing diabetes.

Algorithm Hybrid Diabetes Predictive method	
INPUT	Pima Indian diabetes dataset
OUTPUT	Enhance outcomes predictability and interoperability with reduced ill-posed problems, over-fitting problems, and class imbalance problems
1.	Read the Pima Indian Diabetic Dataset
2.	Apply data pre-processing for improving quality of dataset with better effectiveness
2.1.	Missing value identification and imputation with their respective median value
2.2	Outliers identification and replacement of those values above the 95th and below the 5th percentiles with their respective percentile values
2.3	Random over sampling to reduce class imbalance problem
2.4	Z-score normalization for rescaling numeric columns in a dataset
3.	Apply Lasso regression for feature selection and regularization as well as reduce ill-posed problems or over-fitting problems
4.	Apply artificial neural network for classification.
5.	Model performance is measured using evaluation measures.

4. Results and Discussion

This section evaluates the efficiency of the proposed hybrid diabetes prediction models using the Pima Indian Diabetes dataset. The suggested framework comprises many strategies for improving result predictability and interoperability with reduced ill-posed issues, over-fitting problems, and class imbalance problems for detecting diabetes mellitus utilizing data mining techniques. These methods are median-based missing value imputation, outliers' identification based on the interquartile range, and replacement of those values above the 95th and below the 5th percentiles with their respective percentile values, Random oversampling-based class balancing, Z-score normalization based rescaling numeric columns in a dataset, Lasso regression-based feature selection, and regularization, reduces ill-posed problems or over-fitting problems, and ANN-based classification. A confusion matrix is used to describe the ANN result. Figure 4 depicts the confusion matrix derived from the suggested hybrid diabetes prediction model. The proposed framework properly detects 186 data instances and wrongly determines 14 data instances. Furthermore, the accuracy, precision, recall, f-measure, and ROC parameters are used to examine the performance of the proposed hybrid diabetes prediction models. The confusion matrix is used to determine all of these characteristics. While the ROC curve can be obtained by plotting a graph between the true positive rate and the false positive rate. This is a graphical representation used to represent the capability of our prediction system [19], as shown in figure 5. Whereas in the confusion matrix, True Negative (TN) data points correctly represents the total number of non-diabetes patient, True Positive (TP) data points correctly refers the total number of diabetes patient, False Positive (FP) data points refers the total number of non-diabetes patient classified as diabetes patient, and False Negative (FN) data points shows the number of diabetes patient classified as non-diabetes patient [20]. Table 6 shows the experimental observation of the proposed hybrid model are compared with four machine learning algorithms-based models, while Figure 3 shows its graphical representation, respectively.

In terms of all performance measures, the results show that the Lasso-ANN achieved 93% accuracy with 0.929 precision, 0.929 recall, 0.929 f-measure, and 0.930 ROC. The proposed model is compared to other implemented models based on supervised learning named SVM, k-NN, Naïve Bayes, and J48 algorithms. It found that Lasso-ANN surpassed the other methods, as shown in table 6. As a result, our proposed diabetes prediction method can more precisely identify diabetic patients. Figure 3 represents performance evaluation among the implemented diabetes predictive model, and Figure 4 shows the confusion matrix of the lasso-ANN model. Figure 5 shows the ROC curve of the Lasso-ANN model.

Table 5. Description and formulation of performance metrics.

Metrics	Description	Formula
Accuracy	Calculate the algorithm's accuracy, which is the percentage of properly categorised cases among all occurrences.	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Classifier accuracy is measured by precision, which is defined as the rate of correct predictions.	$\frac{TP}{TP+FP}$
Recall	Used to evaluate classifier completeness	$\frac{TP}{TP+FN}$
F-Measure	Weighted average of precision and recall	$2 \times \frac{Precision \times Recall}{Precision + Recall}$
ROC	Receiver Operating Characteristic Curve	-

Table 6. Experimental observations of five machine learning algorithms.

Algorithms	Accuracy	Precision	Recall	F-Measure	ROC
SVM	90.25	0.951	0.907	0.928	0.880
Naive Bayes	82.46	0.908	0.832	0.868	0.820
k-NN	87.01	0.931	0.879	0.904	0.864
J48	87.01	0.958	0.850	0.901	0.882
Lasso-ANN	**93.00**	**0.929**	**0.929**	**0.929**	**0.930**

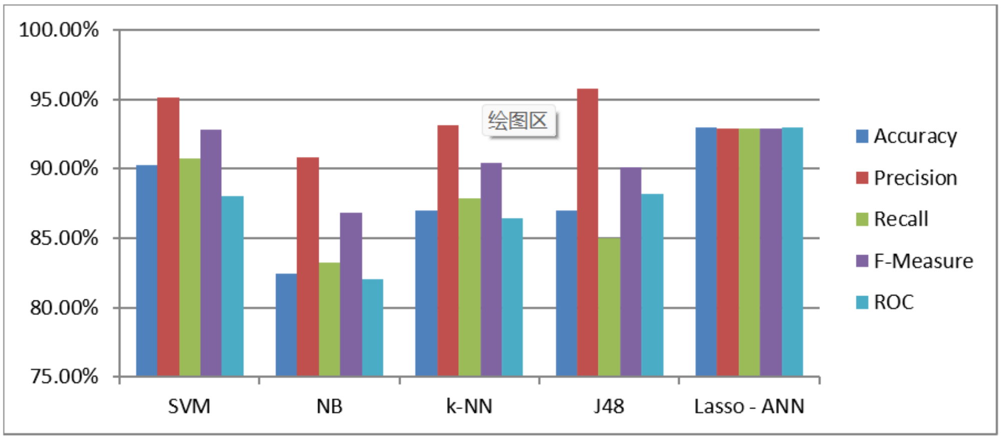

Fig.3. Performance evaluation result of the predictive models.

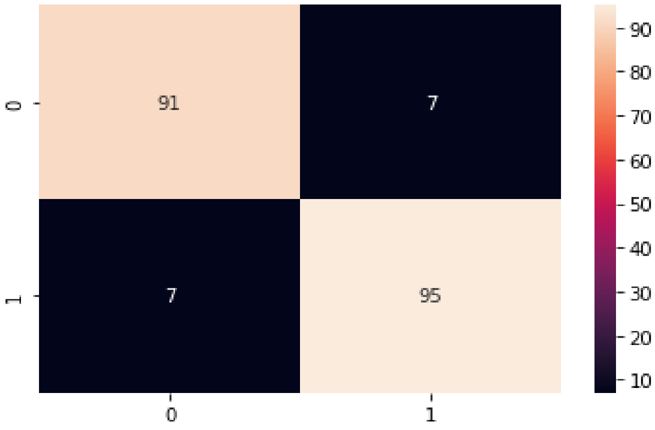

Fig.4. Confusion matrix of Lasso – ANN model.

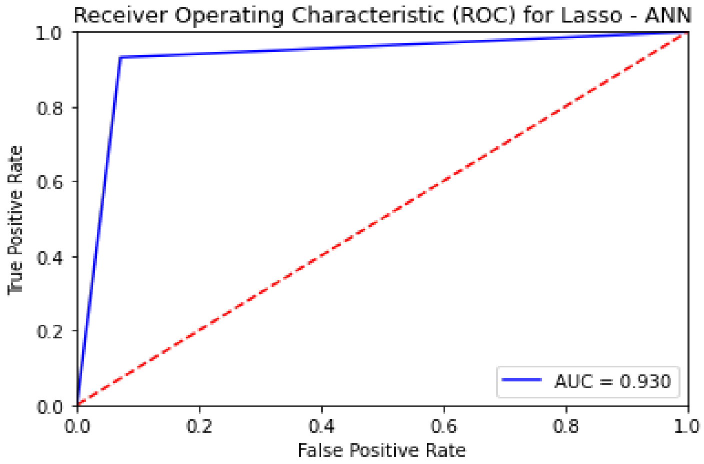

Fig.5. ROC Curve of Lasso – ANN model

Table 7. Comparison of our experimental observations to prior studies.

Sr. No.	Year	Machine Learning Algorithms	Dataset	Performance Metrics				
				Accuracy	Precision	Recall	F-Measure	ROC
1	2020 [21]	SVM	PID	79.15	82.30	49.55	61.87	72.02
		KNN		87.61	88.29	73.45	80.19	84.20
		NB		77.34	71.11	56.63	63.05	72.35
2	2018 [22]	NB	PID	76.30	0.759	0.763	0.760	0.819
		SVM		65.10	0.424	0.651	0.513	0.500
3	2019 [23]	KNN	PID	79.22	NA	NA	NA	NA
		SVM		81.81	NA	NA	NA	NA
4	2018 [24]	SVM	PID	72.73	NA	NA	NA	NA
5	2019 [25]	J48	PID	87.89	NA	NA	NA	NA
		KNN		82.94	NA	NA	NA	NA
		NB		88.41	NA	NA	NA	NA
6	Our Study	SVM	PID	90.25	0.951	0.907	0.928	0.880
		Naive Bayes		82.46	0.908	0.832	0.868	0.820
		k-NN		87.01	0.931	0.879	0.904	0.864
		J48		87.01	0.958	0.850	0.901	0.882
		Lasso-ANN		**93.00**	**0.929**	**0.929**	**0.929**	**0.930**

We also compared our results to those of five prior research papers. It demonstrated that our method is delivering the desired results. The PID Dataset, which contains the same characteristics (described in table 2), was also utilised in these studies [12]. The collected results of these investigations are shown in Table 7.

Table 7 compares the performances of our proposed methods and demonstrates that Lasso-ANN outperforms all other algorithms. The accuracy was 93%, with 0.929 precision, 0.929 recall, 0.929 f-measure, and 0.930 ROC. In the table 7, the accuracy of k-NN is 87.61% [21], Na ïve Bayes is 76.30% [22], SVM is 81.81% [23], SVM is 72.73% [24], and Na ïve Bayes is 88.41% [25] which is less than the proposed model Lasso-ANN. The comparative analysis shows that the proposed model outperformed the state-of-the-art technique in terms of accuracy, as shown in table 7.

5. Conclusion

Our research study's key contribution was the development of a hybrid prediction model that used machine-learning methods, Lasso regression, and artificial neural networks to identify people at high risk of acquiring diabetes. We used the Lasso regression approach for variable selection and regularization, as well as ANN for classification, on the PIMA indian diabetes dataset. Using data mining approaches, we dealt with the issue of ill-posed problems, over-fitting difficulties, and class imbalance problems in the diagnosis of diabetes mellitus. Our models have a high sensitivity and a great capacity to recognize people with DM.

In recent years, diabetes disease has been more prevalent, contributing to a rise in the number of deaths in poor and middle-income countries. Early detection and preventive strategies for diabetes can help to avert disease and save human lives. The major goal of this study is to find the most accurate model and accuracy for predicting diabetes patients. The accuracy of machine learning algorithms that were used in the preceding five years was investigated. As a result, the authors developed a classifier model that incorporates Lasso regression and artificial neural networks. The Pima Indian Diabetes dataset has been taken for experiment and compared with other proposed methodologies; it was found that the performance of the Lasso-ANN achieved the highest accuracy (93%). In the future, the performance may be enhanced by incorporating many pre-processing steps and using more datasets with hybrid-ensemble deep learning models or algorithms.

Reference

[1] World Health Organization. Classification of Diabetes Mellitus. WHO. Geneva: 2019.

[2] Han Cho N. IDF Diabetes Atlas. 2019th ed. 2019.

[3] Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express 2018;4:243–6. https://doi.org/10.1016/J.ICTE.2018.10.005.

[4] Awotunde JB, Ayo FE, Jimoh RG, Ogundokun RO, Matiluko OE, Oladipo ID, et al. Prediction and classification of diabetes mellitus using genomic data. Intell IoT Syst Pers Heal Care 2021:235–92. https://doi.org/10.1016/B978-0-12-821187-8.00009-5.

[5] Yuvaraj N, SriPreethaa KR. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Clust Comput 2017 221 2017;22:1–9. https://doi.org/10.1007/S10586-017-1532-X.

[6] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express 2021. https://doi.org/10.1016/J.ICTE.2021.02.004.

[7] Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng 2021;2:40–6. https://doi.org/10.1016/j.ijcce.2021.01.001.

[8] Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 2017;97:120–7. https://doi.org/10.1016/j.ijmedinf.2016.09.014.

[9] Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J Diabetes Metab Disord 2020;19:391–403. https://doi.org/10.1007/s40200-020-00520-5.

[10] Le TM, Vo TM, Pham TN, Dao SVT. A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic. IEEE Access 2021;9:7869–84. https://doi.org/10.1109/ACCESS.2020.3047942.

[11] Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data 2019;6. https://doi.org/10.1186/s40537-019-0175-6.

[12] Pima Indians Diabetes Database | Kaggle n.d. https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed July 26, 2021).

[13] Subhash AR, Ashwin Kumar UM. Accuracy of classification algorithms for diabetes prediction. Int J Eng Adv Technol 2019;8:230–4.

[14] Soni M, Varma DS. Diabetes Prediction using Machine Learning Techniques. Int J Eng Res Technol 2020;9.

[15] Abdulaziz M, Al-Motairy B, Al-Ghamdi M, Al-Qahtani N. Building a Personalized Fitness Recommendation Application based on Sequential Information. IJACSA) Int J Adv Comput Sci Appl n.d.;12:2021.

[16] ML Studio (classic): Normalize Data - Azure | Microsoft Docs n.d. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalize-data (accessed July 26, 2021).

[17] Malav A, Kadam K. A hybrid approach for Heart Disease Prediction using Artificial Neural Network and K-means. Int J Pure Appl Math 2018;118:103–9.

[18] Mahboob Alam T, Iqbal MA, Ali Y, Wahab A, Ijaz S, Imtiaz Baig T, et al. A model for early prediction of diabetes. Informatics Med Unlocked 2019;16. https://doi.org/10.1016/j.imu.2019.100204.

[19] Meyer-Baese A, Schmid V. Statistical and Syntactic Pattern Recognition. Pattern Recognit Signal Anal Med Imaging 2014:151–96. https://doi.org/10.1016/B978-0-12-409545-8.00006-6.

[20] Kulkarni A, Chong D, Batarseh FA. Foundations of data imbalance and solutions for a data democracy. Data Democr Nexus Artif Intell Softw Dev Knowl Eng 2020:83–106. https://doi.org/10.1016/B978-0-12-818366-3.00005-8.

[21] Jashwanth Reddy D, Mounika B, Sindhu S, Pranayteja Reddy T, Sagar Reddy N, Jyothsna Sri G, et al. Predictive machine learning model for early detection and analysis of diabetes. Mater Today Proc 2020. https://doi.org/10.1016/j.matpr.2020.09.522.

[22] Sisodia D, Sisodia DS. Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci., vol. 132, Elsevier B.V.; 2018, p. 1578–85. https://doi.org/10.1016/j.procs.2018.05.122.

[23] Khurana G, Kumar PA. Improving Accuracy for Diabetes Mellitus Prediction Using Data Pre-Processing and Various New Learning Models. Int J Sci Res Sci Technol 2019:502–15. https://doi.org/10.32628/IJSRST196294.

[24] Mirzajani SS, salimi siamak. Prediction and Diagnosis of Diabetes by Using Data Mining Techniques. Avicenna J Med Biochem 2018;6:3–7. https://doi.org/10.15171/ajmb.2018.02.

[25] Joshi R, Alehegn M. Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. Int Res J Eng Technol 2017;04.

Authors' Profiles

Yogendra Singh received M.Tech. degree in Computer Technology from University of Allahabad, Prayagraj (Allahabad), Uttar Pradesh, India in 2018. He is currently pursuing Ph.D. degree with University of Allahabad, Prayagraj (Allahabad), Uttar Pradesh, India.

Mahendra Tiwari received his Ph.D. in Computer Science from Uttar Pradesh Rajarshi Tandon Open University, Uttar Pradesh, India. His research interests include Data Mining and Soft Computing. He is currently Assistant Professor in computer science at the department of electronics and communication, University of Allahabad, Prayagraj (Allahabad), Uttar Pradesh, India.

How to cite this paper: Yogendra Singh, Mahendra Tiwari, "A Novel Hybrid Approach for Detection of Type-2 Diabetes in Women Using Lasso Regression and Artificial Neural Network", International Journal of Intelligent Systems and Applications(IJISA), Vol.14, No.4, pp.11-20, 2022. DOI:10.5815/ijisa.2022.04.02