World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:18, No:12, 2024

An Approach for the Prediction of Diabetes via Relief Feature Selection

Nebi Gedik

***Abstract*—**One of the most common chronic diseases in the world, diabetes is brought on by insufficient insulin production by the pancreas or by inefficient insulin utilization by the body. The disease is linked to the interplay of lifestyle, behavioral and medical circumstances, demographics, and genetic risk factors. Early disease detection is crucial for helping medical professionals with diagnosis or prognosis as well as for creating a successful preventative strategy. Machine learning techniques are utilized for this purpose in order to identify diabetes from medical records. Finding the characteristics or features that provide the best prediction of classification for diabetes detection is the aim of this study. The performance of each feature is compared using the linear discriminant analysis and k-nearest neighbor classifiers. The feature that yields the best classification results has been determined.

***Keywords*—**Diabetes, relief feature selection, k-nearest neighbor classifiers, lenear discriminant analysis.

I. Introduction

HEREDITARY and chronic illnesses pose a global hazard to public health. One of these conditions is diabetes, a long-term metabolic illness that affects people and raises blood pressure and blood sugar levels. Diabetes can seriously harm (or even kill) key organs, including the kidneys, heart, and nerves, if it is not identified early and treated properly [1]. The disease can have long-term consequences, including cardiovascular dysfunction, cerebral vascular disease, joint failure, sexual dysfunction, and renal and retinal failure [2]. Diabetes continues to be a serious health issue in most nations, despite advancements in technology and medicine. Regardless of income level, 578 million people worldwide are expected to have diabetes by 2030 and 700 million by 2045 [3], [4]. Effective diabetes therapy depends heavily on early diagnosis, as it does for most chronic illnesses. Studies reveal that the probability of recovering from the illness is directly correlated with the early diagnosis [3]. Once more, early diagnosis of conditions like diabetes is now feasible and can be completed more quickly and affordably because of technological advancements. AI and machine learning technologies are making it feasible to extract reliable information from medical data, which includes disease diagnosis [5]. These automated support systems help lower the incidence of disease and enhance quality of life while giving medical personnel more time and attention.

Various studies are being conducted to detect diabetes from medical data using machine learning models. Katarya et al. [6] employed ML approaches using the Python data deployment tool to predict the outcomes and risk factors in the Pima Indian diabetes dataset. Six machine learning models were used in the study: gradient boosting, Support Vector Machines (SVM), Decision Tree (DT), Logistic Regression (LR), k-Nearest Neighbor (k-NN), and Random Forest (RF). Through investigation, it was shown that RF outperforms other methods with a recall of 76%, an area under the ROC curve (AUC) of 83, an accuracy of 84%, a precision of 83%, and an f1-score of 86%. A framework consisting of a DT, a SVM, and the authors' original twice-growth deep learning network was presented by Olisah et al. [7]. They used polynomial regression and Pearson correlation to choose important features. Next, grid-search hyperparameter tweaking and stratified k-fold cross-validation were used to assess the model's performance. Using two datasets—the Pima Indian Diabetes data and the early-stage diabetes risk prediction—the previous researchers [8] developed a model for diabetes prediction. Several models, including Naïve Bayes, k-NN, DT, LR, RF, SVM, AdaBoost classifier, Gradient classifier, and extra tree classifier, are used to detect diabetes. With an accuracy of 86% for the Pima Indian dataset, the Super Learner Classifier model is the most accurate of the several suggested techniques. For the purpose of early diabetes prediction, a fine-tuned multilayer perceptron (MLP) was proposed by [9] considering the good performance of MLP in healthcare and its effectiveness in predicting diabetes complications. There are three primary steps in the suggested approach. Firstly, the sigmoid activation function is applied to the layer outputs and the initial weights to determine the output layer output. Second, at the hidden layer level, the inaccuracy of all hidden units is determined. Lastly, to minimize network mistakes, all network weights are modified backwards. The effectiveness of the fine-tuned MLP was compared to that of several different machine learning techniques, including K-mean, fuzzy c-means clustering (FCM), ANN, and convolutional neural network (CNN). Out of all the comparing approaches, the fine-tuned MLP showed the best performance for early diabetes prediction. To develop and trace predictive machine learning models, [10] employed LR, SVM, boot gradient methods, Naive Bayes, RFs, closest K neighbors. The best prediction models were found to be learning-based models from RF forecasts and booting gradients, with predictive capabilities of 86.28% and 86.29%, respectively. Tasin et al. [11] focus on insulin characteristic prediction with a semi-supervised model with high gradient boosting in order to address the imbalance between classes. Methods like Adaptive Synthetic (ADASYN) sampling and the Synthetic Minority

Nebi Gedik is with the University of Health Science, Institute of Health Science, Turkiye (e-mail: nebi.gedik@sbu.edu.tr).

Over-sampling Technique (SMOTE) are applied. With an XGBoost classifier, the proposed method achieved the maximum accuracy of 81%, an AUC of 0.84, and an F1 score of 0.81, demonstrating the effectiveness of gradient boosting in handling skewed data distributions. Regarding performance analysis, Lyngdoh et al. [12] investigated five supervised machine learning algorithms to estimate the risk of diabetes. The accuracy of the KNN classifier increased to up to 76% when all known risk indicators were consistently included. This result emphasizes how important thorough feature selection is for creating prediction models. A diabetes detection method involving ensemble learning approach was proposed by Rupapara et al. [13]. They used a publicly accessible dataset (Pima India) in their investigation. They also used eight different machine learning models to evaluate the system's efficacy. They used Chi-2, Principal Component Analysis (PCA), and the original features in a variety of experiments. The findings show that Chi-2 characteristics perform better than other features, and the suggested tri-ensemble model achieved an 85% accuracy rate in predicting diabetes cases. The goal of Butt et al.'s [14] machine-learning strategy was to identify and classify diabetes in its early stages. Additionally, they suggested a fictitious Internet of Things (IoT)-based system for tracking blood glucose (BG) levels in both diabetic and healthy people. The classification of diabetes required the application of MLP, LR, and RF. They used linear regression, moving average (MA), and long short-term memory (LSTM) approaches for predictive analysis. The MLP model produced an 86.08% classification accuracy, while the LSTM model achieved an 87.26% prediction accuracy, according to the results.

Diabetes prediction is carried out in [15] by identifying and utilizing significant features, as well as by examining the connections among various features. For diabetes diagnosis, clustering, prediction, and association rule mining are employed, while the principal component analysis approach is favored for identifying significant features. Based on the results, it can be concluded that diabetes is strongly correlated with both body mass index (BMI) and glucose levels measured using the apriori approach. K-means clustering, RF, and artificial neural networks are the classifiers used to predict diabetes. The artificial neural network approach yielded the highest accuracy value of 75.7%. The goal of another study [16] is to develop a model that can most accurately forecast the possibility of diabetes. Three machine learning classification algorithms— DT, SVM, and Naive Bayes —are included in the model that is being presented in order to identify diabetes early on. The Pima Indians Diabetes Database (PIDD), which is part of the UCI machine learning repository, is used to create the model. Accuracy, precision, F-measure, and recall metrics are employed in the performance investigation of the model's outputs. Both correctly and incorrectly identified cases are used to determine accuracy. When compared to the other algorithms, the classification accuracy achieved with Naive Bayes yields the best performance result, with the highest value of 76.30%. Additionally, Receiver Operating Characteristic analysis is used to analyze the results. Additionally, a model [17] is developed to predict if a patient has diabetes based on particular diagnostic parameters in the dataset. The study looks into several methods to increase accuracy and performance. The National Institute of Diabetes and Digestive and Kidney Diseases' PIDD and Vanderbilt's research of rural African Americans in Virginia are served as the two datasets utilized to evaluate the model. For feature selection, two distinct approaches are determined. LR and ensemble approaches, which are considered to improve performance by producing better predictions than a single model, are employed for classification. Compared to ensemble strategies like maximum voting and stacking, the best classification accuracy achieved with maximum voting was approximately 78% for dataset 1 and 93% for dataset 2. In light of the claim that diabetes may be managed if it is identified early, a model aiming more accurate early diabetes diagnosis in a patient is offered in [18] using a variety of machine learning approaches. The dataset in the study is subjected to ensemble techniques and classification algorithms. RF, LR, DTs, SVM, gradient boosting, and k-NN are the classifiers. The results show that, in comparison to other techniques, the RF strategy yields a higher accuracy value of 77%. Models for diabetes detection, which include four classification algorithms, are shown in comparison with two distinct datasets in [19]. These algorithms are RF, SVM, Naive Bayes, DTs (supervised learning algorithms) and k-means (an unsupervised learning algorithm). One of the datasets is obtained from Frankfurt Hospital in Germany and the other is PIDD, provided by the UCI machine learning repository. The performance evaluations of the classification algorithms are carried out with accuracy, F1 score, and recall methodologies. The most successful result obtained with the PIDD dataset is 83.1% with the SVM algorithm. The Synthetic Minority Oversampling, Genetic Algorithm, and DT techniques are used in [20] to classify diabetes using the PIDD dataset. There are four steps in the suggested prediction model. Preprocessing is the initial step, which involves identifying outliers and processing missing values. The second step is the feature selection, where the most beneficial features are identified using a genetic algorithm and correlation. Training the suggested model is the third step. The classification accuracy, classification error, precision, recall, measure, and Area_ Under_ROC metrics are used to assess the outcomes in the final step. The suggested procedure yields an accuracy value of 82.1256%. The greatest outcomes in terms of accuracy, classification error, precision, recall, measure, and Area_Under_ROC are 82.1256%, 17.8744%, 0.8070%, 0.8598, 0.8326, and 0.8511, respectively.

This paper presents an ML model that classifies diabetes data using the relief feature selection approach in conjunction with the k-NN and Linear Discriminant Analysis (LDA) classification methods. Using the relief technique, several feature sets are generated, and each feature set's effectiveness is evaluated using two distinct classifiers.

II. Materials and Method

A. Relief Algorithm

The Relief algorithm was first introduced by Kira and Rendell [21], [22] as a simple, fast, and effective way to weigh attributes. It was motivated by instance-based learning. The Relief algorithm outputs a weight for each characteristic, ranging from -1 to 1. Features with higher weights are considered more selective.

By choosing an example from the data, the closest neighbor examples from the same class (nearest hit) and the opposite class (nearest miss) are located. When a class change coincides with an attribute value change, the attribute is weighted according to the hypothesis that the attribute change might be the reason for the class change. In contrast, an attribute's weight will decrease if its value changes, but the class stays the same because the attribute change has no effect on the class. This method is carried out either for every sample in the data or for a random selection of the samples to update the weight of the attribute. Subsequently, an average is calculated for each weight update, yielding a final weight that lies between [-1, 1]. Relief uses a probabilistic approach to estimate the attribute weight. According to the nearest hit and miss given, respectively, it is proportionate to the difference between two conditional probabilities, or the chance that the attribute value would change [23].

B. Data Set

The model in this study is validated using the PIDD [24]. This dataset contains information that can be used to determine if a patient has diabetes or not, including whether the patient has received a diabetes diagnosis. The number of pregnancies, age, skinfold thickness, blood pressure, insulin, BMI, diabetes pedigree function, and label data are all included in the record. Of the 768 women over 21 who have been observed, 500 do not have diabetes, and 268 have.

C. Methodology

The model comes with an application that integrates data preprocessing, feature extraction, and classification—the three fundamental elements of machine learning. Preprocessing is the first step in the procedure once the dataset is acquired. The original data are now examined for any mistakes or missing data. The relief process is then used four times for feature selection, resulting in the creation of four feature sets of varying sizes. The feature sets with varying sizes are referred to as k1, k2, k3, and k4. The success rates of the k-NN and LDA classifiers are assessed when each feature set is submitted to them. The method's flow chart is displayed in Fig. 1.

Following the implementation of machine learning algorithms, a few tools are employed to assess the quality of the classification process in a number of areas. In machine learning research, these resources are gathered under the heading of performance evaluation metrics. Numerous indicators are employed in the research to evaluate and disclose the effectiveness of specific algorithmic components. Using distinct performance evaluation metrics or metric sets for various machine learning challenges has become essential if they are to be classified. Several standard measures were employed in this study to conduct a comparative analysis and gather useful data regarding the method's performance. These criteria used to assess the classification performance include f-1 score, recall, accuracy and precision.

D. k-NN Algorithm

A fundamental machine learning technique for categorization tasks is KNN. The algorithm is initially created by Fix and Hodges in the early 1950s [25] and modified version of it is presented by Cover and Hart [26]. In order to predict new data points based on their similarity to existing data points in a feature space, the k-NN method needs access to a feature space that contains training data points. The algorithm calculates the distances between an unknown data point and the closest 'k' training data points, which are the number of data points chosen from the training dataset. It designates the new point to the closest majority class. The Euclidean distance (1) is typically the metric selected by the algorithm to compute the distance. The class determination model with the k-NN algorithm is illustrated in Fig. 2 for four training samples. Metrics of measurement like the Manhattan (2) and Minkowski (3) distances are also utilized for distance measurement [27]. The main steps of the k-NN algorithm:

1. Ascertain what k is worth.
2. Determine the separation between each training sample and the test sample.
3. Sort the distance to get the k-neighbors that are nearest.
4. Establish which category the closest neighbors fall into.
5. As the new data object's predictive value, use the simple majority of the nearby neighbor's category.

$$D = \sqrt{\sum(x - y)^2} \tag{1}$$

$$D = \sum|x - y| \tag{2}$$

$$D = \left(\sum|x - y|^p\right)^{1/p} \tag{3}$$

where x and y stand for the point coordinates and d for the distance between two points.

The number of nearest neighbors used in the k-NN algorithm has a significant impact on classification accuracy. Therefore, until the algorithm determines the most ideal k value for the dataset we are working on, it should be tried with various values. This is the k-NN algorithm's time-consuming and negatively expressed function, particularly when dealing with big datasets. It is fairly easy to implement and interpret, and the training time is minimal when the k value is set appropriately.

E. LDA

In 1936, Ronald Fisher [28] presented Linear Discriminant Analysis, a linear transformation approach for binary classification and dimensionality reduction. Projecting data into a reduced dimensional space that maximizes the inter-class variance and decreases the intra-class variation is the primary objective of this technique [29].

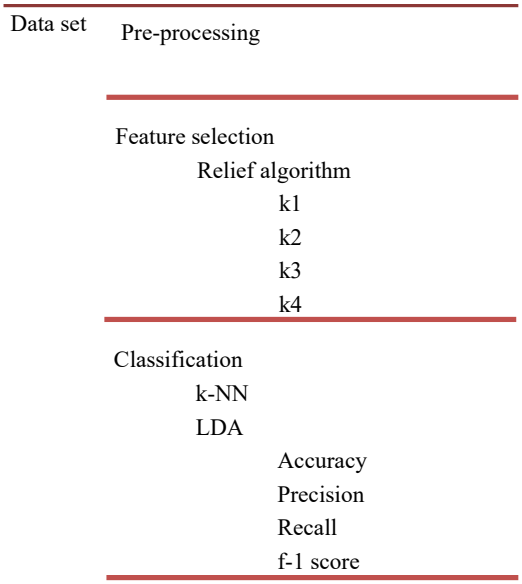


Fig. 1 The method's flow chart

LDA's basic concept is to project a high-dimensional space onto a line, with the goal of minimizing within-class variance and maximizing between-class variance (Fig. 3). The linear discriminant determines a weight vector w that maximizes the Fisher criterion, which is as follows, in order to determine this projection [30]:

$$J(w) = \frac{(w^T\mu_1 - w^T\mu_2)^2}{\sum(y_1 - w^T\mu_1)^2 + \sum(y_2 - w^T\mu_2)^2} \quad (4)$$

The variances are represented by $\sum(y_1 - w^T\mu_1)^2 and \sum(y_2 - w^T\mu_2)^2$, while the means of classes 1 and 2 are denoted by μ_1 and μ_2. We assume that the anticipated samples are represented as $y_1 = w^T x_1$ and $y_2 = w^T x_2$. The Fisher criterion can be expressed as a function of w using these equations as follows [24]:

$$J(w) = \frac{w^T S_B w}{w^T S_w w} \quad (5)$$

where S_w quantifies the within-class scattering and S_B quantifies the distance between the mean values of two classes. Then, using the prior equation, the generalized eigenvalue problem can be solved to determine its maximum [24]:

$$S_B w = \lambda S_w w \quad (6)$$

In this case, w stands for the set of eigenvectors, and λ for the eigenvalues. The eigenvectors are sorted from largest to smallest based on their eigenvalues to form the weight matrix W, which represents the new area to which the data is projected.

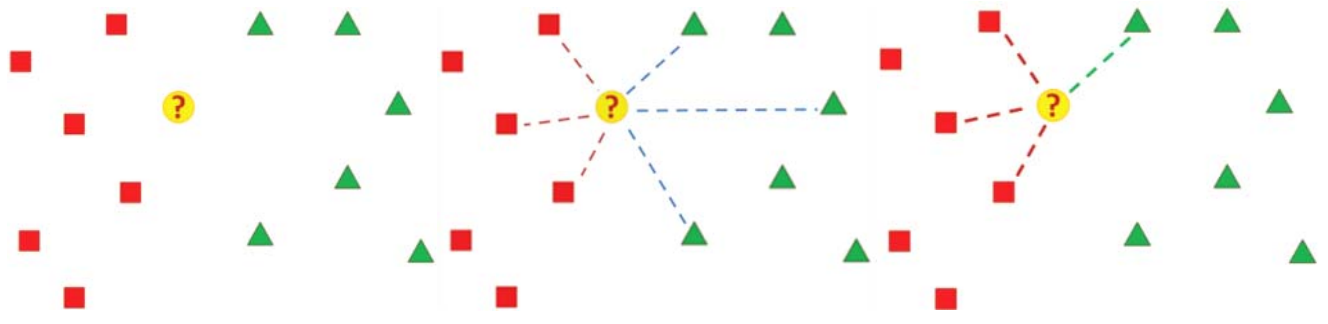
Finding linear combinations of independent variables that permit the identification of classes is the goal of the classification technique known as LDA. It entails recognizing characteristics exclusive to particular classes. The LDA model makes the assumption that each class's data have identical covariances and are normally distributed. The model relies on locating a plane surface with dimensions of n-1 in an n-dimensional space [31].

In Fig. 3, LDA's basic concept is to project a high-dimensional space onto a line. a) A sample of the dataset is represented by each point in the picture. Class information is shown by the red and blue points. When both classes are projected onto the y-axis, the distribution of class data is displayed in the figure. Although the classes have a lot of overlaps and are challenging to successfully separate, this projection adds some separability to the distribution of classes. b) The projection axis with much higher separability as measured by LDA is depicted in the figure. The greatest classification rate is thus attained by optimizing the distance between the class distributions.

F. Accuracy

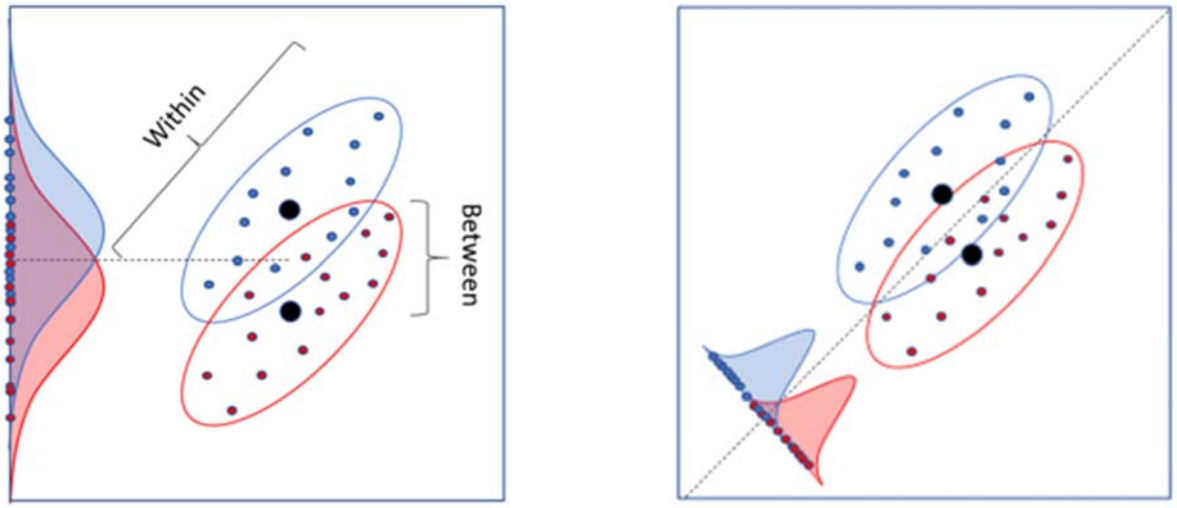
It is the main or most widely used assessment statistic for assessing an algorithm's performance. Its definition is the proportion of accurately categorized data items to all observations (7). It may not be the best performance metric in some circumstances, such as when the dataset is uneven.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (7)$$



(a) (b) (c)

Fig. 2 k-NN process steps: (a) two-class data and the data to be categorized; (b) measuring the distance between the samples; (c) choosing the nearest k values for the choice; for this example, k = 4 [32]



(a) (b)

Fig. 3 LDA's basic concept is to project a high-dimensional space onto a line: (a) The projection with little separability, (b) The projection with much higher separability [33]

G.Precision

It displays the proportion of discovered data items that are relevant or the proportion of observations that the algorithm anticipated to be positive that are in fact positive. The precision can be calculated by dividing the total number of false positives and true positives by the number of true positives (8):

$$Precision = \frac{TP}{(TP+FP)} \quad (8)$$

H.Recall

It gives an indication of the proportion of truly positive observations that the algorithm accurately predicted. The recall is the number of true positives divided by the total of true positives and false negatives (9):

$$Recall\ (Sensitivity) = \frac{TP}{(TP+FN)} \quad (9)$$

where FP is a false negative, TN is a true negative, FP is a false positive, and TP is a true positive.

I. f1 Score

The f-score, also called the f-measure, is a metric used to assess an algorithm's performance that combines precision and recall. The f-measure is the harmonic mean of precision and recall (10):

$$f-1 = \frac{2*Precision*Recall}{Precision+Recall} \quad (10)$$

III. Results

80% of the dataset is used for training, while 20% is used for testing. Training and test data are used in the successive procedures described in the preceding section. The relief method generates four feature sets, which are then assessed using k-NN and LDA classifiers, in that order. Tables I and II display the performance values for each feature set and classifier. Figs. 4-7 provide comparative graphs of each performance metric for the two classifiers.

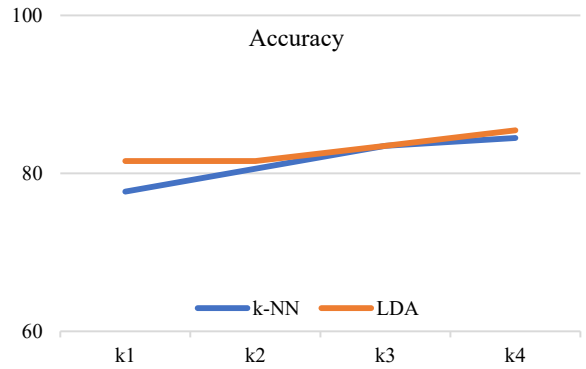


Fig. 4 LDA and k-NN classifier's accuracy scores for each feature set

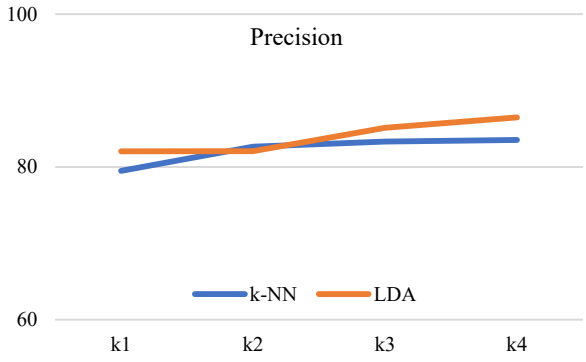


Fig. 5 LDA and k-NN classifier's precision values for each feature set

In the proposed work, an artificial intelligence framework is provided to forecast diabetes illness. A model that classifies patient data as abnormal (not healthy, i.e., having diabetic disease) or normal (healthy) is presented. The suggested approach uses a relaxation algorithm in conjunction with a

feature selection procedure to estimate diabetes risk using the feature data from the PIDD.

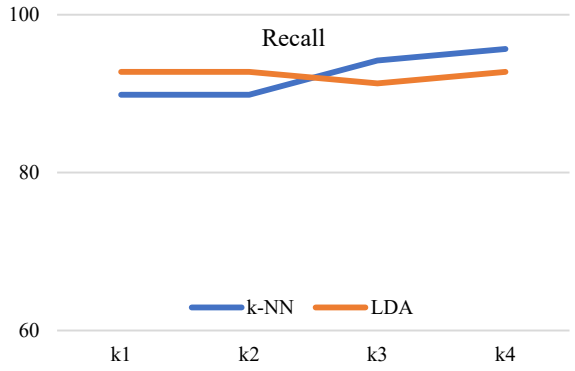


Fig. 6 LDA and k-NN classifier's recall values for every feature set

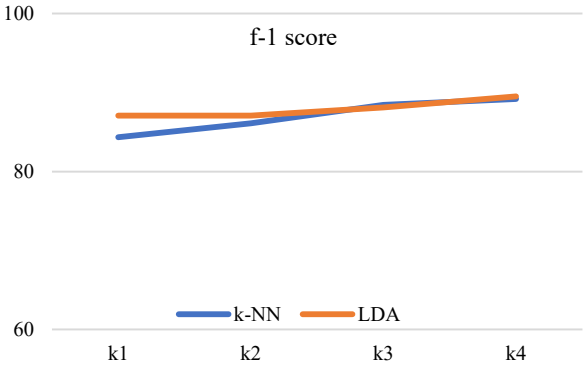


Fig. 7 LDA and k-NN classifier's f-1 score values for each feature set

IV. Conclusion

High blood sugar levels in the body cause diabetes, one of the deadly and chronic diseases that impact many organs and systems. The illness impairs the pancreatic production of insulin and results in malfunctions in the kidneys, heart, eyes, and neurological system. Appropriate therapy depends on early detection; hence, it is critical to design and develop efficient diagnostic instruments. Machine learning algorithms are utilized in the medical field to diagnose and detect problems in patient records.

TABLE I
CLASSIFICATION RESULTS FROM RELIEF + K-NN APPROACH

Relief feature package	k value for k-NN	Accuracy	Precision	Recall	f-1 score
k1	9	77.68	79.49	89.85	84.35
k2	13	80.58	82.67	89.85	86.11
k3	43	83.49	83.33	94.20	88.43
k4	25	84.47	83.54	95.65	89.19

TABLE II
CLASSIFICATION RESULTS FROM RELIEF + LDA APPROACH

Relief feature package	Accuracy	Precision	Recall	f-1 score
k1	81.55	82.05	92.75	87.07
k2	81.55	82.08	92.75	87.07
k3	83.49	85.13	91.30	88.11
k4	85.44	86.49	92.75	89.51

TABLE III
COMPARISON WITH EARLIER STUDIES

	Method	Accuracy
Alam et al. [28]	ANN	75.70
Sisodia et al. [29]	NB	76.30
Tigga et al. [30]	LR	75.32
Larabi-Marie-Sainte et al. [31]	Reptree	74.48
Rajendra et al. [32]	Voting	77.83
Alluri et al. [33]	xgb	80.00
Edeh et al. [34]	SVM	83.1
Azad et al. [35]	PMSGD	80.70
Barik et al. [36]	XG Boost	78.26
Rupapara et al. [13]	LTC (ensemble)	85.00
Proposed	Relief + LDA	85.44

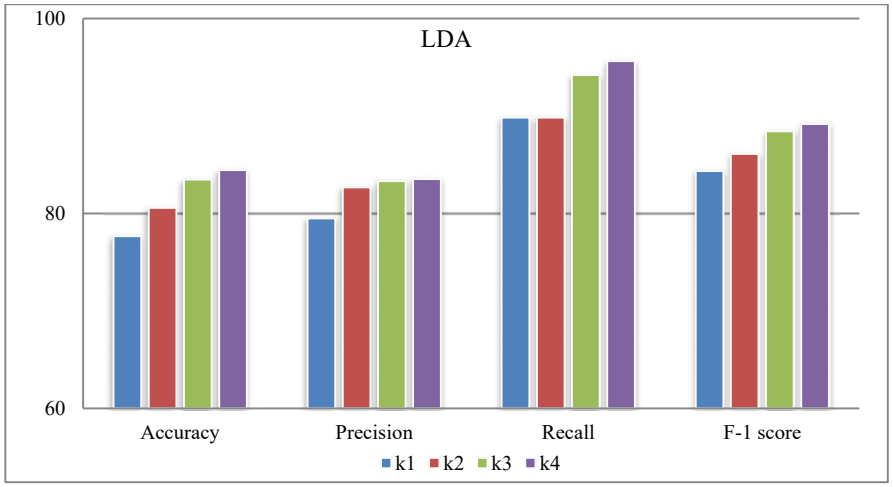


Fig. 6 Comparison of all performance metric values of the LDA classifier according to the generated feature sets

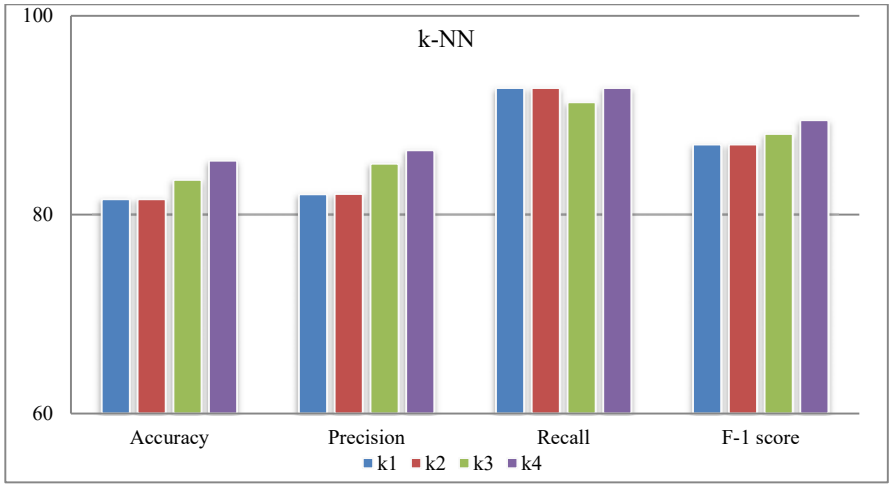


Fig. 7 Comparison of all performance metric values of the k-NN classifier according to the generated feature sets

In the proposed work, an artificial intelligence framework is provided to forecast diabetes illness. A model that classifies patient data as abnormal (not healthy, i.e., having diabetic disease) or normal (healthy) is presented. The suggested approach uses a relief algorithm in conjunction with a feature selection procedure to estimate diabetes risk using the feature data from the PIDD. Using the relief method, four feature sets are generated from the raw data set. Then, they are assessed using k-NN and LDA classifiers comparatively. The most successful result is obtained with the combination of relief k4 and LDA.

REFERENCES

[1] Z. Punthakee, R. Goldenberg, P. Katz, "Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome," *Can. J. diabetes*, vol.42, pp. 10–15, 2018.

[2] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, S. Nalluri, "Genetic algorithm-based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," In *2017 international conference on computing networking and informatics (ICCNI)*, 2017, pp. 1–5.

[3] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, ... & IDF Diabetes Atlas Committee, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas," *Diabetes research and clinical practice*, vol. 157, pp.107843, 2019.

[4] International Diabetes Federation, Facts & figures, (accessed Aug. 27, 2024). Available online: https://www.who.int/health-topics/diabetes#tab=tab_1.

[5] J. Chaki, S. T. Ganesh, S. K. Cidham, & S. A. Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp.3204-3225, 2022.

[6] R. Katarya, S. Jain, "Comparison of different machine learning models for diabetes detection," In: *Proc. 2020 IEEE int. Conf. Adv. Dev. Electr. Electron. Eng. ICADEE 2020. Institute of Electrical and Electronics Engineers Inc.*; 2020, pp. 1-5.

[7] C. C. Olisah, L. Smith, & M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine*, vol. 220, pp.106773, 2022.

[8] S. Saxena, D. Mohapatra, S. Padhee, & G. K. Sahoo, "Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms," *Evolutionary Intelligence*, pp.1-17, 2023.

[9] S. S. Sivasankari, J. Surendiran, N. Yuvaraj, M. Ramkumar, C. N. Ravi, & R. G. Vidhya, "Classification of diabetes using multilayer perceptron," In *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, April 2022, pp. 1-5. IEEE.

[10] L. J. Muhammad, E. A. Algehyne, & S. S. Usman, "Predictive supervised machine learning models for diabetes mellitus," *SN Computer Science*, vol. 1, no. 5, p.240, 2020.

[11] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technol. Lett.* vol. 10, pp.1–10, 2023.

[12] C. Lyngdoh, N. A. Choudhury, & S. Moulik, "Diabetes disease prediction using machine learning algorithms," in *Proceedings 2020 IEEE-EMBS Conf. on Biomedical Engineering and Sciences (IECBES)* (Langkawi Island, Malaysia), 2021, pp. 517–521.

[13] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, "Chi-square and PCA based feature selection for diabetes detection with ensemble classifier," *Intell. Autom. Soft Comput.*, vol. 36, no. 2, pp. 1931–1949, 2023.

[14] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, & H. H. R. Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," *Journal of healthcare engineering*, vol. 2021, no. 1, pp. 9930985, 2021.

[15] K. Kira, L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm", In *proceedings of the tenth national conference on Artificial intelligence*, vol. 2, 1992a, pp. 129–134.

[16] K. Kira, L. A. Rendell, "A practical approach to feature selection," In: *Proceedings of the Ninth International Workshop on Machine Learning*, 1992b, pp. 249–256.

[17] S. F. Rosario, and K. Thangadurai. "RELIEF: feature selection approach," *International journal of innovative research and development*, vol. 4., No. 11, pp. 2018-224, 2015.

[18] Pima Indians Diabetes Dataset, (accessed Aug. 27, 2024). Online. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[19] E. Fix, J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *International Statistical Review/ Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238, 1989.

[20] T. M. Cover, P. Hart, "Nearest neighbor pattern classification," *IEEE Trans Inf Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[21] N. Ali, D. Neagu, & P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," *SN Applied Sciences*, vol. 1, pp. 1-15, 2019.

[22] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.

[23] S. Balakrishnama, & A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, pp. 1-8, 1998.

[24] D. Lopez-Bernal, D. Balderas, P. Ponce, & A. Molina, "Education 4.0: teaching the basics of KNN, LDA and simple perceptron algorithms for binary classification problems," *Future Internet*, vol. 13, no. 8, pp. 193, 2021.

[25] P. Xanthopoulos, P. M. Pardalos, T. B.Trafalis, "Linear Discriminant Analysis," In: *Robust Data Mining. Springer Briefs in Optimization.*

Springer, New York, NY, 2013
[26] A. Mucherino et al. "K-nearest neighbor classification," *Data mining in agriculture*, pp. 83-106, 2009.
[27] How Dimension Reduction works (accessed Oct. 18, 2024) https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-dimension-reduction-works.htm
[28] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, & Z. Abbas, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, pp. 100204, 2019.
[29] D. Sisodia, D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578-85, 2018.
[30] N. P. Tigga, S. Garg, "Predicting type 2 diabetes using logistic regression," *In Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems: MCCS*, pp. 491-500, Springer Singapore, 2019.
[31] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, T. Saba, "Current techniques for diabetes prediction: review and case study," *Applied Sciences*, vol. 9, no. 21, pp. 4604, 2019.
[32] P. Rajendra, S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, pp.100032, 2021.
[33] R. P. Alluri, R. Hemavathy, "Diabetes Prediction Using Ensemble Techniques," *International Journal of Applied Engineering Research*, vol. 16, no. 5, pp. 410-5, 2021.
[34] M. O. Edeh, O. I. Khalaf, C. A. Tavera, S. Tayeb, S. Ghouali, G. M. Abdulsahib, N. E. Richard-Nnabu, A. Louni, "A Classification Algorithm-Based Hybrid Diabetes Prediction Model," *Front. Public Health*, vol. 10, pp. 829519, 2022.
[35] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *In Multimedia Systems, Springer Science and Business Media Deutschland GmbH*, Aug. 2022, pp. 1289–1307.
[36] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques," In *Smart Innovation, Systems and Technologies, Springer Science and Business Media Deutschland GmbH*, 2021, pp. 399–409.

Nebi Gedik received his B.S. degree in Electrical and Electronics Engineering from Fırat University in 2001, his PhD degrees in Electrical and Electronics Engineering from Karadeniz Technical University in 2013, and his MSc degree in 2005 from Atatürk University. He is now an Associate Professor at the University of Health Science. His research interests include medical image and signal processing, pattern recognition and machine learning.