A genetic algorithm-based feature selection approach for diabetes prediction

Kirti Kangra[1], Jaswinder Singh[2]
[1]Department of Computer Science and Engineering, Research Scholar of Guru Jambheshwar University of Science and Technology, Hisar, India
[2]Department of Computer Science and Engineering, Faculty of Guru Jambheshwar University of Science and Technology, Hisar, India

Article Info

Article history:

Received Jul 28, 2023
Revised Oct 16, 2023
Accepted Nov 15, 2023

Keywords:

Chronic disease
Feature selection
Genetic algorithm
Health-care
Machine learning
Optimisation

ABSTRACT

Genetic algorithms have emerged as a powerful optimization technique for feature selection due to their ability to search through a vast feature space efficiently. This study discusses the importance of feature selection for prediction in healthcare and prominently focuses on diabetes mellitus. Feature selection is essential for improving the performance of prediction models, by finding significant features and removing unnecessary among them. The study aims to identify the most informative subset of features. Diabetes is a chronic metabolic disorder that poses significant health challenges worldwide. For the experiment, two datasets related to diabetes were downloaded from Kaggle and the results of both (datasets) with and without feature selection using the genetic algorithm were compared. Machine learning classifiers and genetic algorithms were combined to increase the precision of diabetes risk prediction. In the preprocessing phase, feature selection, machine learning classifiers, and performance metrics methods were applied to make this study feasible. The results of the experiment showed that genetic algorithm + logistic regression i.e., 80% (accuracy) works better for PIMA diabetes, and for Germany diabetes dataset genetic algorithm + random forest and genetic algorithm + K-Nearest Neighbor i.e., 98.5% performed better than other chosen classifiers. The researchers can better comprehend the importance of feature selection in healthcare through this study.

Corresponding Author:

Kirti Kangra
Department of Computer Science and Engineering Guru Jambheshwar University of Science and Technology
Hisar, Haryana, India
Email: kirtikangra98@gmail.com

1. INTRODUCTION

Feature selection is a crucial step in machine learning (ML) and data analysis, which involves selecting the most pertinent and instructive features (variables or attributes) from a broader set of potential features [1]. The goal of feature selection is to improve the performance of an ML model by reducing dimensionality, mitigating the risk of overfitting, speeding up the training process, and providing an optimal solution using Optimisation strategies. It can also help in improving model interpretability and reducing noise. Some common methods of optimization strategies for feature selection are:

Filter methods: These methods assess the relevance of each feature independently of the others and select features based on statistical measures [2]. Some common filter methods include correlation, chi-squared test, and variance thresholding. Wrapper methods: These methods involve training and evaluating the ML model with different subsets of features to determine which combination yields the best performance

[3]. Common wrapper methods include: Forward selection, backward elimination and recursive feature elimination (RFE). Embedded methods: Some ML algorithms have built-in feature selection mechanisms. For example L1 regularization (Lasso), tree-based methods, dimensionality reduction techniques, domain knowledge, feature importance from tree-based models, and sequential feature selection.

Metaheuristic algorithms: These are optimisation techniques used to find solutions to complex problems by exploring the search space efficiently. In the context of feature selection, metaheuristic algorithms can be employed to search for an optimal or near-optimal subset of features [4]. These algorithms aim to find the best subset of features by evaluating different combinations based on a fitness or objective function such as genetic algorithms, particle swarm optimisation, or simulated annealing.

Hybrid methods: Combine multiple feature selection techniques to take advantage of their strengths and mitigate their weaknesses [5]. For example, using a filter method to preselect a subset of relevant features and then applying a wrapper method for fine-tuning. The type of data being used, the problem being addressed, and the available computational resources all influence the feature selection. It's often a crucial part of the feature engineering process in ML projects as it can significantly impact model performance and efficiency.

This study primarily focuses on the genetic algorithm (GA) for feature selection to predict diabetes mellitus with the help of different ML classifiers. The GA [6], a search heuristic, is based on Charles Darwin's idea of natural evolution that replicates the process of natural selection in which the fittest individuals are chosen for procreation to develop offspring for the future generation. GA is part of metaheuristic algorithm that is used in feature selection to optimize the solutions to different computer science problems. The key components of a metaheuristic algorithm are intensity and dispersion. To effectively address the real situation, a balance between these components is essential. Metaheuristic algorithms fall into two main categories in computer science: single-solution and population-based metaheuristic algorithms. GA comes under population-based metaheuristic algorithm. It generally incorporates biologically inspired operators like mutation, crossover, and selection to optimize and solve search problems with high-quality solutions. It is frequently employed in problem-solving, research, and ML. The correct sequence of operators (discussed below) is required to be followed while implementing GA (shown in Figure 1 [7]):

- Initialization/Population: In the first step of the GA, the initial population is produced at random for each unique solution. The population is determined by the type of problem, which may have a number of solutions. For the subsequent steps, some encoding schemes must be implemented at this step.
- Selection: The next phase in GA is the reproductive phase. It randomly selects chromosomes from a population, based on an objective function. The objective functions are used to select individuals through survival of the fittest [8].
- Crossover: To produce a new chromosome or offspring, two chromosomes are fused. This procedure is used to develop a new offspring after selection.
- Mutation: It alters the value of one or more genes on the chromosome. This operator randomly flips few chromosome bits [9].
- Fitness Function/Objective Function: It examines various chromosomes to determine which is optimal. This study used the fitness function noted:

$$F(t) = eval(z) \tag{1}$$

Where $z = \frac{a+b}{a+b+c+d}$

a = observation is positive & predicted to be positive

b = observation is negative & predicted to be negative

c = observation is negative & predicted to be positive

d = observation is positive & predicted to be negative

There are various encoding schemes for the chromosome representation (see Figure 1.). In this study Binary encoding scheme is used in which chromosomes are encoded as binary strings and have two potential gene variants, 0 and 1. It is assumed that chromosomes are points in the solution space. These are handled by repeatedly replacing its population with genetic operators. Nowadays, GA is used to solve real-world problems emanating from a variety of disciplines, including economics, medicine, politics, management, and engineering. This study focuses on the use of GA in the healthcare sector.

The world is currently dealing with a number of chronic diseases, including diabetes, cancer, tuberculosis, and heart disease. It is essential to find these diseases early on. These diseases must be endured for a very long time by the sufferer and are spreading more widely every day. To control these diseases, more study is needed. Among them, for this article, we have chosen Diabetes chronic disease. GA and diabetes are the topics of this research.

Predicting and categorizing diabetes mellitus is one of the most difficult tasks in biomedical sciences [10]. Diabetes is one of the top causes of death in developing nations. High diabetes prevalence rates in people aged 20 to 79 were reported for China, India, and Pakistan in 2021. With 0.6 million deaths, India ranks third in the world due to its sizable population and high rate of diabetes. Diabetes is one of the global health calamities with the fastest rate of growth in the twenty-first century, according to the tenth edition of the International Diabetes Federation Atlas. Due to inadequate insulin, glucose levels in people with diabetes continue to rise. A research report from the International Diabetes Federation projects that by 2040, there will be 642 million cases of diabetes worldwide [11]. So, to reduce this number machine learning optimisation techniques are required for better results.

The Key objectives of this paper is to provide an overview of genetic algorithm. Experiment and compare the values of different evaluation parameters using with and without genetic algorithm to demonstrate the importance of feature selection. This paper is divided into sections. The first section of the paper provides an overview of GA and diabetes. The second section discusses Literature related to GA. The rest of the sections discuss the methodology used and experimental results.

Problem statement diabetes can be diagnosed as a binary classification problem that divides all experiment subjects into two groups: those who have diabetes and those who do not. In the article by Kangra and Singh [12] perform an experiment on conventional machine learning classifiers. We have compared different ML classifiers for diabetes prediction but the results are not satisfactory. To enhance the prediction rate, this study will use feature selection. It plays a significant part in disease prediction by removing pointless work and shrinking the dataset. Today, diabetes has developed into a significant issue that requires effective treatment. But using conventional ML classifiers it cannot be predicted in a precise way. To get more precise results this article will use GA for feature selection with conventional ML classifier to predict diabetes.

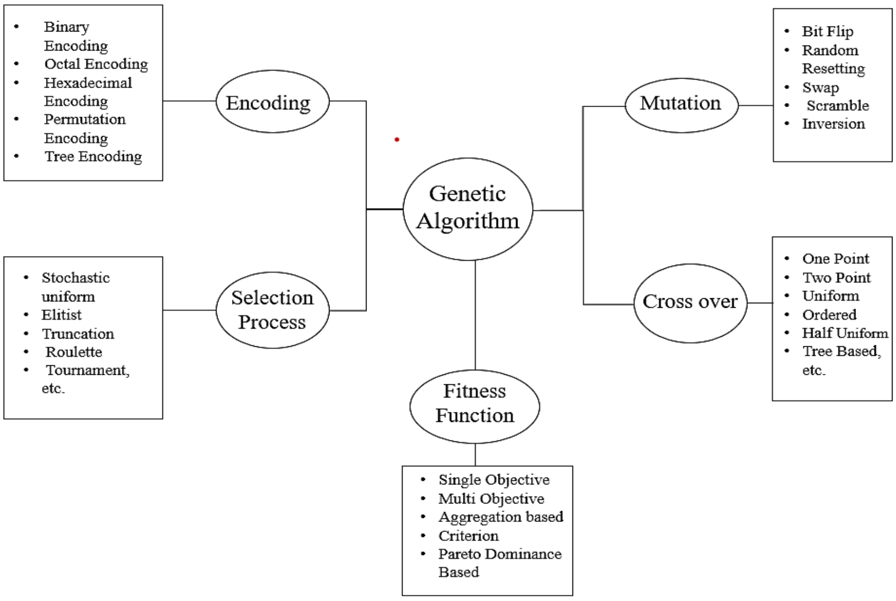

Figure 1.Taxonomy of genetic algorithm

2. REVIEW OF LITERATURE

This section highlights pertinent research on feature selection using GA and classification algorithms that have been used to identify diabetes and other medical conditions including heart disease. To identify diabetes, a number of feature selection methods have been used. Table 1 which was created after analysing the studies described below helps in the selection of classifiers for experimental analysis.

In this study, Patil *et al.* suggested a stacking-based non deterministic sorting genetic algorithm (NSGA-II) technique for type 2 diabetes prediction [13]. On-dominated sorting GA was utilised in NSGA-II. This strategy made use of two diabetes-related datasets. Utilising Matlab experiment was conducted. Comparisons between the proposed NSGA-II stacking methodology and the boosting, bagging, random forest (RF), and random subspace methods were designed. K-Nearest neigbour (KNN) outperforms decision trees when used as a stacking combiner. A novel diabetes detection technique was presented by Domínguez *et al.* [14] employing design of experiment, GA, and multi-layer perceptron (MLP). GA was used to optimise the parameters in the MLP. The MLP serves as a model within GA to provide the fitness evaluation of the solutions. The diabetes statistics were gathered from the "Sylhet Diabetes Hospital in Bangladesh". The proposed model's accuracy rate was 98%. Arukonda and Cheruku [15] designed a medical support system for diabetes prediction. In this research, the researcher combined feature selection algorithms with ML classifiers. Feature selection was performed by the Akaike information criterion and GA. Six well-known classifier algorithms, including support vector machine (SVM), RF, KNN, gradient boosting, extra trees, and naive bayes (NB), were coupled with these methodologies. The dataset was produced using patient records from the general hospital "Centro Médico Siglo XXI" in Mexico. S. Arukonda and R. Cheruku [15] their study they had discussed different chronic diseases i.e. Diabetes, Heart, Kidney, and Breast Cancer. Datasets were split in the ratio of 90:10. 10% of test data was further split using 5-fold cross-validation. The bootstrapped technique was used to apply logistic regression (LR), SVM, decision tree (DT), and KNN basis learners to create 20 different base learners. GA was used to find the best ensemble learner. PIMA diabetes dataset (PIDD), kidney, heart, and breast cancer dataset showed 90.9%, 96.05%, 97.56%, and 98.08% accuracy rates for the proposed model. E1-Shafiey *et al.* [16] discussed heart disease. The researcher proposed a hybrid model consisting of GA, particle swarm optimization (PSO), and RF. For the experiment Python was used on the PIDD [17]. The result achieved by the model was 95.60%. Togatorop *et al*. [18] proposed stacked generalization GA to predict heart disease. PID and Diabetes 130-US hospital datasets were used to check the validity of the model. The result of the proposed model was 98.8 and 99.01%. Researchers in Tan *et al*. [19] proposed an optimisation model to predict heart disease. In this GA and RF were used for the prediction. The result of the anticipated model was 85.83%. Rajagopal *et al*. [20] proposed a model to predict diabetes. In this research, the researcher used “Qingdao desensitization physical examination data from 1 January 2017 to 31 December 2019”. GA, DT, CNN, SVM, and KNN were used. The suggested model produced a result of 98.71%. Nagarajan *et al.* [21] stat ed a customized hybrid model of artificial neural network (ANN) and GA for diabetes prediction. With an accuracy rate of 80%, the mentioned customised hybrid model and its supporting decision-making algorithm were applied to PIDD obtained from the University of California Irvine (UCI) ML Repository. Ashri *et al.* [22] proposed a system for diagnosing cardiovascular disease. The objective of this project was to design a hybrid genetic-based crow search algorithm (GCSA) for feature selection and classification using deep convolution neural networks”. The accuracy of the GCSA model was 95.34% for extracted features and 88.78% for original features. El-Shafiey *et al.* [23] introduced “hybrid classifiers using the ensembled model with majority voting” technique to boost prediction for cardiovascular disease. The dataset was acquired from UCI. LR, SVM, RF, DT, and KNN were used as ML classifiers. The experiment was performed using Python. The result of the proposed model was 98.18%.

Li *et al.* [24] introduced GA-RF based heart disease model. The dataset was taken from UCI. The experiment was performed using Python. The experimental results show that the proposed technique predicted heart disease with 95.6% accuracy on the Cleveland dataset. Dinesh and Prabha [25] proposed a method that had three steps: preprocessing, feature selection, and classification to predict diabetes. “Harmony search algorithm, GA, and PSO algorithms were examined with K-means for feature selection”. The diabetes dataset was classified using KNN. The proposed method had 91.65% accuracy.To anticipate diabetes, Rani *et al*. [26] proposed a hybrid model. In the proposed work, the features were transformed using Kernel principal component analysis (KPCA). SVM was used for classification, and GA to select features. To experiment, Python was used. The presented method had 97.3% accuracy rate. A hybrid decision support system was put forth by Alharan *et al.* [27] in as a tool for the early diagnosis of cardiovascular disease. GA and Synthetic Minority Oversampling Technique were used in the preprocessing phase. NB, LR, SVM, and RF were used as classification models. Among them, RF performed better with an accuracy of 86.6%. To obtain the dataset UCI was used and for the experiment, Python was used. In this study, Dweekat and Lam [28] proposed a diabetes diagnosis system by analysing two different diabetes datasets, namely “PIDD and Dr. John Schorling's data”. Python was used for the experiment. For feature selection, linear discriminant analysis and GA methods were used. For classification, RF, logistic model tree, and JRip algorithms were used. The datasets' accuracy was 90.89% and 91.44%, respectively.

ML methods can help with the early identification of diabetes by looking at related research. But there are also some problems with these investigations. The associated research uncovers the following knowledge gap: i) A few studies failed to use parameter measurements to display their findings. Accuracy is crucial, but performance evaluation also takes into account other factors, such as error rates. As a result, using more prominent features to minimise calculations, GA may be used as a potential tool to make predictions more efficiently. It is possible to design more effective solutions for the healthcare industry by integrating GA with other optimisation techniques. This will aid practitioners in making more accurate decisions in the healthcare industry.

Table 1.Classifier count

ML Classifiers	References	Count
DT	[13], [15], [17], [22]	4
SVM	[14], [15], [17], [19], [22], [25], [26]	7
KNN	[13]–[15], [19], [22], [24]	6
LR	[15], [22], [26]	3
NB	[14], [17], [26]	3
RF	[13], [14], [16], [17], [18], [22], [23], [26], [27]	9
ANN	[20]	1
MLP	[28], [17]	2

3. METHODOLOGY

The essential phases used in this research are shown in Figure 2. There are four phases to this methodology. In the first phase, Preprocessing was used to generate a properly organised dataset. In the second phase, the optimum features were chosen using GA, with the help of an objective function (1). In the third phase, six ML classifiers were used for classification. Finally, the results of the classifier were evaluated using the performance metrics.

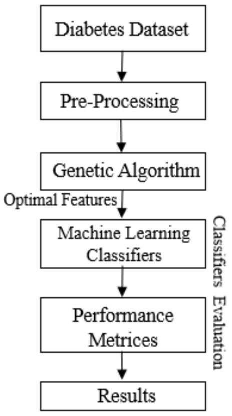

Figure 2. Methodology

3.1. Datasets description

The pima Indian diabetes (PID) [29] and the Germany diabetes [30] datasets were used to evaluate the utility of the feature-selection approach. Both databases are often used to forecast diabetes and are freely accessible. The PID dataset, which was used to evaluate whether a patient had diabetes and was from the "National Institute of diabetes and digestive and kidney diseases (sNIDDK)", was acquired from the UCI ML repository. The PID dataset contains data on 768 females and eight attributes. The Hospital Frankfurt in Frankfurt, Germany provided the second dataset. Eight features and 2,000 cases are included in the dataset.

3.2. Pre-processing phase

In order to use data in an ML model, it must first be cleaned and formatted. This process is known as data pre-processing. It is correct to conclude that this is the first and most crucial step in the process of creating an ML model. There are number of outliers in the dataset like missing values, zeros instead of values, etc these errors need to be tackled. When creating an ML model, having access to adequate, well-structured data is beneficial, but it is not always a given. To make the process more manageable, data preprocessing is divided into four steps: information extraction, data aggregation, data compression, and transformation of data. In this study, zeros were replaced with the mean value.

3.3. Feature selection

The inclusion of numerous additional aspects and features has contributed to a substantial growth in medical data. The majority of features do not influence the outcomes of predictive models but increase calculation time and resource requirements. As a result, in order to achieve high accuracy rates, picking up a limited number of features is required. On the PID and Germany diabetes datasets in this work, the ideal subset of characteristics was selected using GA. The number of features in the datasets was decreased by using the objective function provided in the algorithm. Algorithm opted for the feature selection described:

```
Input: set parameters, Produce P random population solutions with n max number of generations using binary encoding
Output: P(n) the best features
Start
For each individual form i to P do
Evaluate Fitness function
While iteration number < n
Select= SelectBst(i);
If Select then // using tournament selection
 If Cross-over then
    Choose two parents $i_a$ and $i_b$
     Produce off-spring $i_c$= cross-over
Else
     Choose one individual
     Produce off-spring by Mutate($i_c$)
Terminate
    Evaluate the fitness value of $i_c$;
    Replace with least fitness value individual;
```

By following the above steps number of times pregnant, ‘Plasma glucose concentration a 2 hours in an oral glucose tolerance test’; ‘Body mass index (weight in kg/(height in m^2)); Diabetes pedigree function; Age (years); these features among the selected dataset were chosen for PIDD. The same steps were applied to the other diabetes dataset and ‘Glucose’; ‘skinthickness’; ‘Insulin’; ‘BMI’; ‘Diabetes pedigree function’ opted for optimised results.

3.4. Algorithm selection

The dataset and the type of prediction determine which classifier is being used. The NB, LR, SVM, DT, RF, and KNN ML classifiers were chosen from the literature. From Table 1 classifiers whose count>2 were selected for the experiment.

3.5. Software

The experiment was carried out through Python. The Python scikit-learn library contained the ML classifiers (Python offers built-in libraries that can be used to implement different feature selection algorithms). They can be implemented using libraries such as Jupiter Notebook, NumPy, Pandas, and Scikit-learn. For the experiment, Jupiter notebook was run on “AMD Ryzen 5 5500U with radeon graphics and 16 GB RAM under x64 bit Windows 11 operating system”.

4. EXPERIMENT ANALYSIS

For the experiment, this study used PIMA and the Germany diabetes datasets, downloaded from Kaggle. The experiment was conducted in two phases: with and without feature selection using GA. This study shows the comparison between results with and without using GA. The mean was utilised to replace all zeros throughout the pre-processing phase. Table 2 gives details on the parameter values used by GA during the trial. Following preprocessing, 10-fold cross-validation was used to divide the data into training and testing sets. While only 30% of the data was used for testing, 70% of it was used for training.

Accuracy [31], Precision [32], Recall, F-1 score, MCC, Kappa value, AUC [33], MAE, RMSE, RAE, RRSE, and MSTSS parameter metrics and error rates were used in the analysis. Mean absolute error: As computed by averaging the absolute difference over the dataset, it reflects the difference between the actual and expected values [34]. Root mean squared error (RMSE): It is a well-known method for assessing model error when predicting statistical data. RMSE values between 0.0 and 0.5 indicate that the model can make precise predictions about the data. Relative absolute error (RAE): It is a technique for assessing a predictive model’s potency. It contrasts mean mistakes to trivial errors and is expressed as a ratio [35]. Root relative squared error (RRSE) is a key indicator that sheds light on a model's performance. It also has a relative squared error (RSE) variation. Matthews’ correlation coefficient measures the accuracy of categorisation by accounting for true and erroneous positives and negatives. In this, a perfect forecast is represented by a value of 1, an imperfect prediction by a value of 0, and a total difference between the prediction and the observation by a value of 1 [36]. Kappa Value: It is a statistic that measures the agreement between the observed classification and the classification produced by a model while accounting for the possibility of agreement occurring by chance [37].

4.1. Results for diabetes dataset

From Table 3 different classifiers NB, SVM, DT, KNN, LR, and RF showed 77%, 75%, 70%, 74%, 76%, and 77% accuracy scores. Among them, NB and RF perform better with 77%. From Table 4 different classifiers NB, SVM, DT, KNN, LR, and RF showed 70%, 76%, 73%, 78%, 80%, and 78% accuracy scores after feature selection. Among them, LR performs better with 80%. If kappa values (Table 3 and 4) were analysed all the classifiers were not enough strong to predict diabetes for the PID dataset. Figures 3 and 4 depict that among all the selected classifiers NB was the only one whose accuracy and auc decreased after feature selection.

Table 2. Genetic algorithm parameter

Parameter name	Value
Size of Population	100
Number of generations	50
Crossover Probability	0.6
Mutation Probability	0.3
Mutation type	Uniform
Selection type	Tournament
Fitness function	Accuracy
Encoding	Binary

Table 3. PIDD without GA

Classifiers	Accuracy	Precision	F-1 score	Recall	MCC	Kappa Value	Area Under Curve (AUC)
NB	77	71	65	60	0.43	0.43	71
SVM	75	76	63	55	0.45	0.43	67
DT	70	63	59	56	0.43	0.43	65
KNN	74	65	55	47	0.31	0.30	68
LR	76	78	66	57	0.49	0.47	69
RF	77	71	65	60	0.43	0.29	71

Table 4. PIDD with GA

Classifiers	Accuracy	Precision	F-1 score	Recall	MCC	Kappa Value	AUC
NB	70	72	76	77	0.44	0.43	60
SVM	76	77	76	77	0.43	0.42	70
DT	73	73	73	74	0.24	0.23	68
KNN	78	70	60	56	0.40	0.39	72
LR	80	72	62	53	0.46	0.45	72
RF	78	70	65	58	0.44	0.43	73

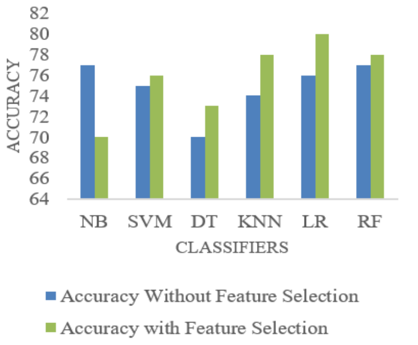

Figure 3. Comparison between accuracy

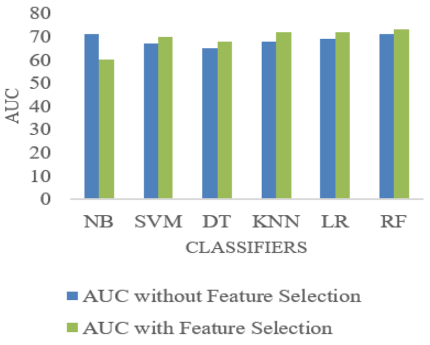

Figure 4. Comparison between AUC

4.2. Results for Germany diabetes dataset

From Table 5 different classifiers NB, SVM, DT, KNN, LR, and RF showed 76.5%, 77%, 94%, 98%, 77%, and 98% accuracy scores. Among them, KNN and RF perform better with 98%. From Table 6 different classifiers NB, SVM, DT, KNN, LR, and RF showed 78.6%, 77.8%, 95%, 98%, 77.5%, and 99% accuracy scores after feature selection. Among them, KNN and RF perform better with 98.5 %. Figures 5 and 6 demonstrate that RF and KNN show almost 1% improvement after using GA. By analysing kappa values (Tables 5 and 6) DT, KNN, and RF are enough strong to predict diabetes for the Germany dataset.

Table 5.Without GA

Classifiers	Accuracy	Precision	F1-score	Recall	MCC	Kappa Value	AUC
NB	76.5	75	62	76	0.47	0.45	82
SVM	77	76	70	77	0.50	0.45	71
DT	94	94	92	94	0.87	0.89	97
KNN	98	98	93	98	0.97	0.97	98
LR	77	77	60	77	0.50	0.45	83
RF	98	98	94	98	0.97	0.98	99

Table 6.With GA

Classifiers	Accuracy	Precision	F1-score	Recall	MCC	Kappa Valuue	AUC
NB	78.6	77	62	77	0.39	0.39	84
SVM	77.8	78	72	78	0.47	0.45	72
DT	95	93	92	95	0.87	0.87	98
KNN	98.5	98	94	98	0.97	0.98	99
LR	77.5	78	64	78	0.51	0.50	84
RF	98.5	98	96	98	0.97	0.98	98

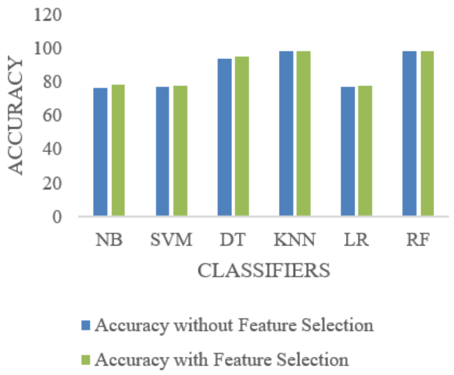

Figure 5. Comparison between accuracy

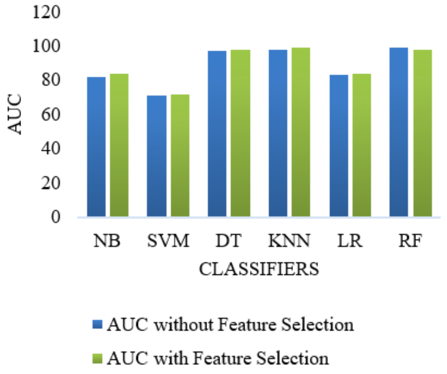

Figure 6. Comparison between AUC

4.3. Error rates of chosen classifiers

For each classifier, the error rates were examined in Tables 7 and 8. For best outcomes, the error rate should be as low as possible. One would frequently choose the classifier with the lowest RAE and RRSE or the lowest MAE and RMSE, depending on their needs. Low RRSE values (in Tables 7 and 8) are desirable for improved classifier predictions. In the PID dataset, RRSE values are very high as compared to the Germany dataset for some classifiers. All classifiers have the same RAE of 0.86, so they perform equally well in terms of RAE for the PID dataset. If both MAE and RMSE are equally important, the study could consider LR as the best choice since it has the lowest values for both. DT and KNN produce low RRSE values for the Germany dataset, making them suitable as classifiers. Given the emphasis on MAE and RMSE as measures of prediction accuracy, the DT classifier appears to be the best performer among all the classifiers in this specific context. Without feature selection error rates were computed in the previous article of ours.

The study used two different datasets for diabetes that showed results with and without feature selection. It provides a clear view that after feature selection accuracy of almost every classifier was increased and also reduced error rates. Therefore, based on the results of the experiment described above, it can be said that feature selection is crucial for improving prediction accuracy.

Table 7. Error rates of PIDD using GA

Classifiers	MAE	RMSE	RAE	RRSE%
NB	0.26	0.51	0.86	51.80
SVM	0.27	0.52	0.86	52.22
DT	0.36	0.60	0.86	59.94
KNN	0.28	0.53	0.86	53.45
LR	0.25	0.50	0.86	50.96
RF	0.26	0.51	0.86	51.80

Table 8. Error rates of Germany using GA

Classifiers	MAE	RMSE	RAE	RRSE%
NB	0.26	0.51	0.77	51.63
SVM	0.27	0.52	0.77	52.12
DT	0.05	0.23	0.77	25.16
KNN	0.23	0.48	0.77	48.47
LR	0.27	0.52	0.77	52.12
RF	0.26	0.51	0.77	51.63

5. CONCLUSION

Thus, GA can be used as an optimisation technique to find the best optimal solution. The GA provides optimal features to predict diabetes through initialisation, selection, crossover, mutation, and replacement. To obtain results the study followed four steps: Pre-processing, feature selection, ML classifiers, and performance metrics. Consequently, after feature extraction LR performs better with an accuracy rate of 80% for PIMA diabetes and 98.5% accuracy rate for the Germany diabetes dataset for RF and KNN. In addition, the study compared the results of both datasets with and without feature selection. It shows that by using the feature selection algorithm results for the selected ML classifiers can be increased. Accuracy, Precision, Recall, F-1 score, MCC, Kappa value, AUC, MAE, RMSE, RAE, and RRSE parameter metrics and error rates were used in the analysis phase. Thus, from this experiment related to diabetes, we can say that GA alone is not sufficient for the accurate prediction of diabetes. There should be some other method that integrates with GA to make its performance much better for other parameters i.e., Kappa value, and MCC. The combining GA with some other optimisation methods is the future concern of this research paper. The datasets utilised here have fewer instances and properties, but the study may use some sizable datasets.

REFERENCES

[1] H. Polat, H. Danaei Mehr, and A. Cetin, “Diagnosis of chronic kidney disease based on support vector machine by feature selection methods,” *Journal of Medical Systems*, vol. 41, no. 4, p. 55, Apr. 2017, doi: 10.1007/s10916-017-0703-x.

[2] R. Parthiban et al., “Prognosis of chronic kidney disease (CKD) using hybrid filter wrapper embedded,” *European Journal of Molecular & Clinical Medicine*, vol. 07, no. 09, pp. 2511–2530, 2021.

[3] M. Manonmani and S. Balakrishnan, “An ensemble feature selection method for prediction of CKD,” *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*, 2020, doi: 10.1109/ICCCI48352.2020.9104137.

[4] V. Kumar, J. K. Chhabra, and D. Kumar, “Parameter adaptive harmony search algorithm for unimodal and multimodal optimization problems,” *Journal of Computational Science*, vol. 5, no. 2, pp. 144–155, Mar. 2014, doi: 10.1016/j.jocs.2013.12.001.

[5] A. Prabha, J. Yadav, A. Rani, and V. Singh, “Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier,” *Computers in Biology and Medicine*, vol. 136, 2021, doi: 10.1016/j.compbiomed.2021.104664.

[6] Z. Michalewicz, “Genetic Algorithms + Data Structures = Evolution Programs 3ed _Michalewicz.PDF,” *3rd ed. Springer Berlin Heidelberg*, 1996.

[7] U. Mehboob, J. Qadir, S. Ali, and A. Vasilakos, “Genetic algorithms in wireless networking: techniques, applications, and issues,” *Soft Computing*, vol. 20, no. 6, pp. 2467–2501, Jun. 2016, doi: 10.1007/s00500-016-2070-9.

[8] A. Ramesh, C. Kambhampati, J. Monson, and P. Drew, “Artificial intelligence in medicine,” *Annals of The Royal College of Surgeons of England*, vol. 86, no. 5, pp. 334–338, Sep. 2004, doi: 10.1308/147870804290.

[9] P. Ghodmare, “A review paper on brief introduction of genetic algorithm,” *International Journal of Science Technology & Engineering*, vol. 4, no. 8, pp. 42–44, 2018.

[10] G. R. Ashisha, X. A. Mary, H. M. Ashif, I. Karthikeyan, and J. Roshan, “Early diabetes prediction with optimal feature selection using ML based prediction framework,” in *2023 4th International Conference on Signal Processing and Communication (ICSPC)*, Mar. 2023, pp. 391–395. doi: 10.1109/ICSPC57692.2023.10125956.

[11] I. D. Federation, “Facts & Figures,” *International Diabetes Federation*, 2023.

[12] K. Kangra and J. Singh, “Comparative analysis of predictive machine learning algorithms for diabetes mellitus,” *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1728–1737, Jun. 2023, doi: 10.11591/eei.v12i3.4412.

[13] R. N. Patil, S. Rawandale, N. Rawandale, U. Rawandale, and S. Patil, “An efficient stacking based NSGA-II approach for predicting type 2 diabetes,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, p. 1015, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1015-1023.

[14] A. García-Domínguez *et al.*, “Diabetes detection models in mexican patients by combining machine learning algorithms and feature selection techniques for clinical and paraclinical attributes: a comparative evaluation,” *Journal of Diabetes Research*, vol. 2023, pp. 1–19, Jun. 2023, doi: 10.1155/2023/9713905.

[15] S. Arukonda and R. Cheruku, “A novel diversity-based ensemble approach with genetic algorithm for effective disease diagnosis,” *Soft Computing*, vol. 27, no. 14, pp. 9907–9926, Jul. 2023, doi: 10.1007/s00500-023-08393-5.

[16] M. G. El-Shafiey, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, “A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest,” *Multimedia Tools and Applications*, vol. 81, no. 13, pp. 18155–18179, May 2022, doi: 10.1007/s11042-022-12425-x.

[17] J. Abdollahi and B. Nouri-Moghaddam, “Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction,” *Iran Journal of Computer Science*, vol. 5, no. 3, pp. 205–220, Sep. 2022, doi: 10.1007/s42044-022-00100-1.

[18] P. R. Togatorop, M. Sianturi, D. Simamora, and D. Silaen, “Optimizing random forest using genetic algorithm for heart disease classification,” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 13, no. 1, p. 60, Aug. 2022, doi: 10.24843/LKJITI.2022.v13.i01.p06.

[19] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, “Early risk prediction of diabetes based on GA-stacking,” *Applied Sciences*, vol. 12, no. 2, p. 632, Jan. 2022, doi: 10.3390/app12020632.

[20] A. Rajagopal, S. Jha, R. Alagarsamy, S. G. Quek, and G. Selvachandran, “A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures,” *Mathematics and Computers in Simulation*, vol. 198, pp. 388–406, Aug. 2022, doi: 10.1016/j.matcom.2022.03.003.

[21] S. M. Nagarajan, V. Muthukumaran, R. Murugesan, R. B. Joseph, M. Meram, and A. Prathik, “Innovative feature selection and classification model for heart disease prediction,” *Journal of Reliable Intelligent Environments*, vol. 8, no. 4, pp. 333–343, Dec. 2022, doi: 10.1007/s40860-021-00152-3.

[22] S. E. A. Ashri, M. M. El-Gayar, and E. M. El-Daydamony, “HDPF: heart disease prediction framework based on hybrid classifiers and genetic algorithm,” *IEEE Access*, vol. 9, pp. 146797–146809, 2021, doi: 10.1109/ACCESS.2021.3122789.

[23] M. G. El-Shafiey, A. Hagag, E.-S. A. El-Dahshan, and M. A. Ismail, “Heart-disease prediction method using random forest and

genetic algorithms," in *2021 International Conference on Electronic Engineering (ICEEM)*, Jul. 2021, pp. 1–6. doi: 10.1109/ICEEM52022.2021.9480625.

[24] X. Li, J. Zhang, and F. Safara, "Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm," *Neural Processing Letters*, vol. 55, no. 1, pp. 153–169, Feb. 2023, doi: 10.1007/s11063-021-10491-0.

[25] M. G. Dinesh and D. Prabha, "Diabetes mellitus prediction system using hybrid KPCA-GA-SVM feature selection techniques," *Journal of Physics: Conference Series*, vol. 1767, no. 1, p. 012001, Feb. 2021, doi: 10.1088/1742-6596/1767/1/012001.

[26] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, Sep. 2021, doi: 10.1007/s40860-021-00133-6.

[27] A. F. H. Alharan, Z. M. Algelal, N. S. Ali, and N. Al-Garaawi, "Improving classification performance for diabetes with linear discriminant analysis and genetic algorithm," in *2021 Palestinian International Conference on Information and Communication Technology (PICICT)*, Sep. 2021, pp. 38–44. doi: 10.1109/PICICT53635.2021.00019.

[28] O. Y. Dweekat and S. S. Lam, "Optimized design of hybrid genetic algorithm with multilayer perceptron to predict patients with diabetes," *Soft Computing*, vol. 27, no. 10, pp. 6205–6222, May 2023, doi: 10.1007/s00500-023-07876-9.

[29] UCI Machine Learning, "Pima Indians diabetes database," 2021, doi: Pima Indians Diabetes Database | Kaggle.

[30] DaSilva John, "diabetes | Kaggle," 2022, [Online]. Available: https://www.kaggle.com/datasets/johndasilva/diabetes?resource=download

[31] N. Cahyani and M. A. Muslim, "Increasing accuracy of C4.5 algorithm by applying discretization and correlation-based feature selection for chronic kidney disease diagnosis," *Journal of Telecommunication, Elctronic and Computer Engineering*, vol. 12, no. 1, pp. 25–32, 2020.

[32] M. Gollapalli *et al.*, "A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM," *Computers in Biology and Medicine*, vol. 147, p. 105757, Aug. 2022, doi: 10.1016/j.compbiomed.2022.105757.

[33] H. Lu, S. Uddin, F. Hajati, M. A. Moni, and M. Khushi, "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," *Applied Intelligence*, vol. 52, no. 3, pp. 2411–2422, Feb. 2022, doi: 10.1007/s10489-021-02533-w.

[34] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[35] R. Naseem *et al.*, "Empirical assessment of machine learning techniques for software requirements risk prediction," *Electronics*, vol. 10, no. 2, p. 168, Jan. 2021, doi: 10.3390/electronics10020168.

[36] L. Ali *et al.*, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," *IEEE Access*, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS.2019.2909969.

[37] K. C. Howlader *et al.*, "Machine learning models for classification and identification of significant attributes to detect type 2 diabetes," *Health Information Science and Systems*, vol. 10, no. 1, p. 2, Feb. 2022, doi: 10.1007/s13755-021-00168-2.

BIOGRAPHIES OF AUTHORS

Kirti Kangra is pursuing her Ph.D. from Guru Jambheshwar University of Science and Technology, Hisar, Haryana in Computer Science and Engineering. She has teaching experience for 2 years. She has completed her B. Tech in Information and Technology from Vaish College of Engineering, Rohtak, and M. Tech in Computer Science and Engineering from the University Institute of Engineering and Technology affiliated with Maharishi Dayanand University, Rohtak, Haryana. Her areas of research include machine learning, data mining, and artificial intelligence. She can be contacted at email: kirtikangra98@gmail.com.

Jaswinder Singh is working as Professor in the Department of Computer Science & Engineering at Guru Jambheshwar University of Science &Technology, Hisar, Haryana. He has teaching experience of more than 20 years and he has published more than 25 research papers in international journals and conferences. He has completed his Ph.D in Computer Science & Engineering from Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonepat, Haryana and completed his M.Tech in Computer Science & Engineering from Kurukshetra University, Kurukshetra, Haryana. His areas of research include machine learning, opinion mining, web information retrieval, search engine optimisation, web mining, information processing, information system, and social network analysis. He can be contacted at email: Jaswinder_singh_2k@rediffmail.com.