Journal of Namibian Studies, 33 S2 (2023): 6781-6798 ISSN: 2197-5523 (online)

Comparison Of Three Classification Methods For Feature Selection In Diabetes Data

Dr. Senada Bushati, Msc. Anxhela Gjecka

Aleksander Moisiu University, Durres, Albania
bushatin@yahoo.com

Abstract

This paper describes a focus on predicting and classifying people suffering from diabetes using 3 classification techniques. It aims to provide a correct diagnosis at the right time to prevent fatal outcomes. The diabetes is a significant global health concern and is one of the leading causes of death worldwide. The paper employs various techniques to predict and classify individuals at risk of diabetes. These techniques include the filter method, wrapper method, and genetic algorithm method. These methods are often used for feature selection and model building in machine learning. The study suggests that while heuristic methods may not be as accurate as classification methods, the results are satisfactory. The AUC (Area Under the Curve) value reached 80% in a hybrid combination of the genetic algorithm (GA) method with the GSA (Gravitational Search Algorithm). The proposed system can easily distinguish between healthy and unhealthy individuals, which is essential for early intervention and treatment. The algorithms used in the study have a running times and memory usage by 98.75%. The combination of different algorithms, such as GA and GSA, can help doctors diagnose sick patients efficiently and on time. The main objective, the paper appears to offer a method for predicting and classifying diabetes, highlighting the importance of early diagnosis and the potential benefits of using a combination of different algorithms to improve efficiency and accuracy. 80% of AUC value suggests that the model's performance is quite promising in distinguishing between healthy and diabetic individuals. However, it's important to note that the effectiveness of such models can vary depending on

the quality and quantity of data used in training and testing.

Keywords: Diabetes prediction, machine learning algorithms, filter and wrapper methods, genetic algorithms, GSA, AUC, classifier.

Introduction

Diabetes is a disease that comes as a result of some metabolic disorders that are caused by high levels of glycemia in the blood. This leads to some other diseases caused by it such as heart disease, stroke, kidney failure, blood vessels, etc. Statistically, it is a disease that has alarmingly increasing cases of people suffering from one of the types of diabetes and also has high mortality rates. Diabetes is categorized into three types: Type I diabetes (T1D), Type II diabetes (T2D), and gestational diabetes (GD) [5]. Therefore, in this study, we will use machine learning techniques to diagnose diabetes but also find the best classifier to more accurately determine this disease. Feature selection is an efficient data preprocessing technique to reduce the dimensionality of the data [6]. It is very important to identify the most important risk factors associated with the disease in medical diagnoses. Identification of relevant features helps to remove unnecessary and redundant attributes from the patient's dataset, achieving optimal and faster results. Classification and prediction are data mining techniques that train data to develop a model and then the discovered model is applied to the test data. To obtain the prediction results, various algorithms have been applied to the clustering of diabetes data and very promising results have been found. There is an utmost need to develop a new technique that can accelerate and simplify the diagnosis process of this disease. Summarizing our work, we organize our analysis along with two research points and formulate the following research questions:

- RQ1. How do these parameters affect the prediction?
- RQ2. How effective the algorithm is for the prediction?
- RQ3. How will SA affect the improvement of model accuracy?

Based on the research questions we assume two hypotheses:

- ✓ H1. The combination of features based on one of the classification methods combined with SA affects the result of the return of the model.

- ✓ H2: The model works best if we use metaheuristic algorithms or other algorithms.

The contributions of this work are:

1. Proposing some classification methods for the most accurate forecast for the classification of sick people and their treatment on time.
2. Development of a metaheuristic method for classification (ranking) and prediction using familiar classifiers.
3. Finding an optimization method or a meta-heuristic optimization method that has as its main goal to achieve a high degree of prediction accuracy.

The result obtained from one method will be combined with the GSA method to achieve a hybrid method. In the last years, have become more prevalent (popular) hybrid methods for diagnosing and predicting chronic diseases. These are divided into two stages: the first stage is 'feature selection', which is used to select a subset of features, and the second stage is used to build models based on the subset created by the first stage [23]. Based on prediction and classification we are going to adopt; we will interpret whether the selected hybrid method provides satisfactory results on the obtained dataset to predict people at risk of diabetes or whether the chosen algorithm is not satisfactory enough for these data [24]. This material begins firstly explaining the impact that machine learning has on creation of health algorithms. We will explain the filtering methodology, the wrapper methodology, and the genetic algorithm as they will be applied to the considered dataset and finally interpret the results obtained from the use of the methods. Thereafter it will be applied Generalized Simulated Annealing to smooth the hyperparameters. We will explain how it behaves and what needs to be improved. This text appears to be an excerpt from a research proposal or paper focused on using machine learning techniques to diagnose diabetes and improve the accuracy of diabetes classification. Let's break down the key points and objectives mentioned in this text:

Related Work

We applied different machine learning algorithms to a dataset. One of these algorithms is Logistic Regression, which achieved an accuracy of 96% in classifying the data. Logistic Regression is often used for binary classification tasks. After evaluating multiple algorithms, it appears that the AdaBoost classifier was chosen as the best model for this specific dataset. AdaBoost is

an ensemble learning method that combines multiple weak learners to create a strong classifier. In this case, it achieved an impressive accuracy of 98.8%. The study involved a comparison between two different datasets. From what we've described, it seems that this new dataset, when used with the chosen machine learning model (AdaBoost), provided better accuracy and precision for predicting diabetes compared to an existing dataset. This suggests that the new dataset and model combination is an improvement. The new model improved both accuracy and precision for diabetes prediction. This is important because accuracy measures the overall correctness of predictions, while precision is a measure of how many of the positive predictions made by the model are correct. High accuracy and precision are generally desirable in medical applications like diabetes prediction to minimize false positives. [9]. Referring to [11], experiments were performed to predict diabetes in Indian Pima women with a particular ML classifier. 768 female patients were considered for this study. Various data extraction tasks were performed to perform a comparative analysis of four different ML classifiers: Naïve Bayes (NB), J48, Logistic Regression (LR), and Random Forest (RF). These models were analyzed with different cross-values of control (K = 5, 10, 15, and 20).

The performance of the machine learning models was evaluated using various metrics, including accuracy, precision, recall, F1 score, and AUC (Area Under the Receiver Operating Characteristic Curve). These are common metrics used to assess the performance of classification models, particularly in the context of medical diagnosis. The preliminary results suggest that all the researched models performed well. This means that the models showed promise in accurately predicting whether a patient has diabetes. It appears that Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF) were identified as the top three models for predicting diabetes. These models likely demonstrated the best trade-off between accuracy and other relevant factors. The study appears to involve the analysis of various patient-related data, including lifestyle, inherited information, and other factors that may contribute to diabetes. The primary goal is early disease identification and treatment. Data mining techniques are used to uncover patterns and insights within this data, which can aid in early diagnosis and improved treatment. The paper suggests the use of preprocessing procedures and data reduction strategies. The main goal is to select a minimal set of significant features that result in the highest classification accuracy when using SVMs. Four two-objective meta-heuristic algorithms are used. These algorithms are likely designed to

optimize two conflicting objectives: maximizing classification accuracy and minimizing the number of selected features. Meta-heuristic algorithms are commonly used for feature selection because they can efficiently explore the feature space and find good solutions. Support vector machines (SVMs) are used as the classification model. SVMs are a popular choice for binary and multi-class classification tasks due to their ability to find an optimal hyperplane that maximizes the margin between different classes. The accuracy values obtained through the hybrid method are 98.2% and 94.6%, respectively. This means that the hybrid approach using the selected features managed to achieve these levels of accuracy in the classification task. These findings demonstrate the potential of the proposed approach in solving classification problems with high accuracy while using a reduced set of features. [12]

The purpose [13] of this research is to choose a powerful machine learning algorithm that can be applied in both diabetes and liver disease prediction. On two independent datasets, diabetes and liver illness, this study examines two machine learning methodologies, SVM (Support Vector Machine) and KNN (nearest K-neighbors) algorithms. It was discovered that a tuned radial SVM approach had the highest accuracy in detecting diabetes and liver disease, with 0.989 accuracies for diabetes detection and 0.91 accuracies for liver disease detection. Medical decision support systems (DSSs) continue to demonstrate their usefulness in providing clinical decision support to physicians and other healthcare professionals. The article proposes a DSS for diabetes prediction based on machine learning (ML) techniques and deep learning approaches. The most extensively used classifiers for the traditional machine learning method are Vector Support Machines (SVM) and Random Forest (RF). In Deep Learning (DL), on the other hand, a fully convolutional neural network (CNN) was utilized to predict and detect diabetes patients. The suggested approach was tested using the public Pima Indians Diabetes database, which included 768 samples with eight attributes each. There were 500 non-diabetic samples and 268 diabetes patients in the study. The overall accuracy obtained using DL, SVM, and RF was 76.81%, 65.38%, and 83.67%. The experimental results show that RF was more effective in predicting diabetes than deep learning and SVM methods. [14]

This work is focused on the classification of diabetic survey data, particularly in the context of imbalanced categories and complex characteristics. The study involves the use of multiple

supervised classifiers, a synthetic data generation technique called SVM-SMOTE, and two-dimensional dimensionality reduction methods (stepwise logistic regression and LASSO).

The performance of the classification models is evaluated using several important metrics, including accuracy, precision, recall, F1-Score, and AUC (Area Under the ROC Curve). These metrics are used to assess how well the models are performing in identifying high-risk diabetic patients and to provide a comprehensive view of their effectiveness. The study's findings indicate that the Random Forest classifier, when paired with the SVM-SMOTE and LASSO feature reduction methods, is particularly effective at identifying high-risk diabetic patients. This suggests that the combination of these methods improves the model's performance in distinguishing high-risk individuals. The research concludes that the combined method proposed in the study can be a valuable tool for early screening of diabetes (DM). Early identification of high-risk diabetic patients is essential for timely intervention and management. (Accuracy = 0.890, Precision = 0.869, Recall = 0.919, F1-Score = 0.893, AUC = 0.948). [15]

The [16] examines the early prognosis of diabetes by referring to data extraction techniques. To test the accuracy of the predictive data extraction algorithms, the dataset gathered 768 instances from the PIMA Indian Diabetes dataset. The five models were tested for accuracy, precision, sensitivity, specifications, and F1 outcome measures using nine input variables and one output variable from the dataset. The purpose study aims to accurate Naive Bayes, logistic regression models, the C5.0 decision tree, and support vector machines (SVMs) in predicting diabetes using common risk factors. The decision tree model (C5.0) had the highest classification accuracy, followed by the logistic regression model, Nave Bayes, and SVM.

The Definition of the Problem

It's alarming to see the significant rise in diabetes diagnoses over the years. The current human lifestyle is the primary cause of diabetes's rise. There are three types of errors in the current medical diagnosing system.

1. The false-negative kind, in which a patient is diabetic in reality but test results show that he or she does not have diabetes.

2. The second type is the false-positive type. In this case, the patient is not diabetic in reality, but test results indicate that he or she is.
3. The unclassifiable category, in which a system is unable to diagnose a specific case. Due to a lack of knowledge extraction from previous data, a given patient may be predicted in an unclassified manner.

This module covers data collection and analysis to order to investigate patterns and trends, which aids in forecasting and evaluating outcomes. The following is a description of the dataset. There are 10000 records in this Diabetes dataset, each with 22 properties. A given patient may be predicted in an unclassified manner due to poor knowledge extraction from previous data.

Diabetes_012, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income

Data Preprocessing

1) Data Cleaning: Before moving on to the application of the hybrid model, the ''Diabetes Prediction'' dataset will go first through the checking phase. First, data cleaning to see if any values do not make sense, or are missing, [20] this replacement of values will be done through the following formula:

$$\underline{x} = \frac{\sum_{j=1}^{n} xj}{n}$$

(1)

2) Data Preparation: At the same time, data customization will be applied the goal of normalization is to adjust the values of results to a fixed range without changing differences in the ranges of values. Normalization aims to give the results a uniform structure. The normalization is defined as:

$$\mathbf{X}_{\text{normalized}} = \frac{X - \mu}{\sigma}$$

(2)

Clasification Techniques

A. Filter Method. The filter method selects those features that have high results while excluding the others. This method is divided into two categories, univariate and multivariate. Multivariate make sure that the correlation between variables is zero or minimal, while univariates disregard any dependencies between variables [19]. We are analyzing multivariate that rely on the correlation between variables. The method we will develop involves selecting features based on correlation. This method selects those variables that have the most characteristics with the target variable and controls and ensures that they have a low correlation between them [19]. Feature selection will be done by observing the correlation between the pairs of variables, the one with the highest value will be the first feature. And so on. The calculation will be carried out using the following formula:

$$M_s = \frac{k\underline{r}_{cf}}{\sqrt{k+k)k-1)\underline{r}_{ff}}}$$

(3)

Where Ms is the heuristic evaluation of a subcategory of features containing such k, while, is the average of the correlation between the features and the targeted variable, and is the average ratio between features [19].

B. Wrapper Method. This method works in the same way as the filter method but has the distinction of using a defined classification algorithm instead of using an independent measure to evaluate subgroups. The wrapper methods give better results than the filter method but have to increase overall if the number of features increases. [19] Subgroup accuracy will also be evaluated by the performance of the calculating time of the algorithm by using the following formula:

$$\mathrm{RED}(\chi)=\frac{1}{\rho(\rho-1)}\sum_{f_i,f_j\epsilon\chi,i>j} \rho(f_i,f_j)|\ )$$

(4)

Where ρ (fi, fj) is the correlation of the person between the features fi and fj. A high value of RED (χ) indicates that the group is strongly related. Therefore, lower values of RED(χ) serve better for feature selection. [18]

C. Genetic Algorithm. The genetic algorithm was proposed by Holland in 1960, using the idea of Darwin's theory. The GA is an algorithm that creates a population based on an initial

group. The basic idea for building this algorithm is based on three fundamental genetic operators which are: crossover, mutation, and selection [20]. The Algorithm works as follows: There are 4 chromosomes X1, X2, X3, and X4. Each chromosome contains 9 genes, and these genes are represented by BITI “1” AND BITI “0”. If the Bit is “1” this indicates that the feature is selected and if the bit is “0”, this feature is not selected [20]. The first genetic operator aims to select possible chromosomes from the population being analyzed. The initial population becomes the parent of the selected population after selecting the features where the bit becomes “1”.

Table 1

	1st	*2nd*	*3rd*	*4th*	*5th*	*6th*	*7th*	*8th*	*9th*
X1	1	0	1	0	1	1	0	0	0
X2	0	0	1	1	1	0	0	1	0
X3	0	1	1	1	0	0	0	1	0
X4	0	1	1	0	0	1	1	0	0

Parent

1	*0*	*1*	*0*	*0*	*1*	*0*	*1*	*0*
0	*1*	*0*	*0*	*1*	*0*	*1*	*1*	*0*

Child

1	*0*	*1*	*0*	*0*	*0*	*1*	*1*	*0*
0	*1*	*0*	*0*	*1*	*1*	*0*	*1*	*0*

Before

1	*0*	*0*	*1*	*1*	*1*	*0*	*1*	*0*

After

1	*0*	*0*	*1*	*1*	*0*	*0*	*1*	*0*

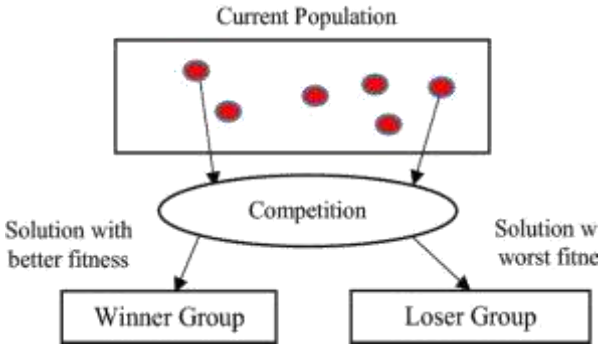
The selection is done randomly based on the probabilistic evaluation. In short, a chromosome with a better value has a

higher probability of becoming the ancestral [20]. This evaluation is carried out using the following formula:

$$P(X_i)=\frac{(1-F(X_i)}{\sum_{i=1}^{N}(1-F(X_i)}$$

(5)

Where N is the I population number, I is the chromosome position in the population, and F (.) is the fitness function. So, the selection scheme is like the one shown in the illustration



D. Generalized simulated annealing. The Sa method aims to define an overall minimum value of each objective function by simulating the process up to final fulfillment [21]. Taking into consideration an objective function f(X) with X = (X1, X2,..., Xn)T, we attempt to determine the possible minimum using Sa. The GSA was proposed by Tsallis and Stariolo [21]. The GSA is constructed from the general entropy as follows:

$$s_q=k\frac{\sum p_t^q}{q-1}$$

(6)

where q is a real number, I am the energy spectrum index and sq determines entropy: [21]

$$s=-k\sum p_i lnp_i$$

(7)

where q 1. We maximize Tsalis entropy as follows:

$$\sum p_i=1$$

(8)

$$\Sigma \quad p_i^q \varepsilon_i$$

(9)

Where εi is the spectrum energy and the distribution probability is evaluated as follows [21]:

$$p_i = \frac{[1-(1-q)\beta\varepsilon_t]^{\frac{1}{1-q}}}{z_q}$$

(10)

where zq is the normalizing constant that ensures that the probability goes to 1. This distribution converges with the Gibbs-Boltzmann distribution where q tends to 1 [21].

Statistical Analyses

The database used for the analysis contains 21 variables and has a size of 5000 observations. The principal conditions of all three methods were to split the database and select only those variables that have an impact on the second generation of the forecast. The parameters selected for the forecast are

('HighBP',

'HighChol',

'CholCheck',

'Stroke',

'HeartDiseaseorAttack',

'Veggies',

'HvyAlcoholConsump',

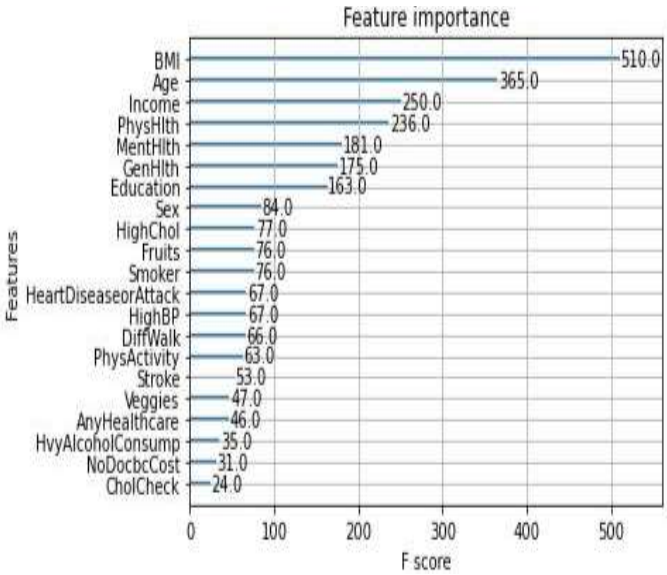
'AnyHealthcare',

'NoDocbcCost',

'GenHlth')

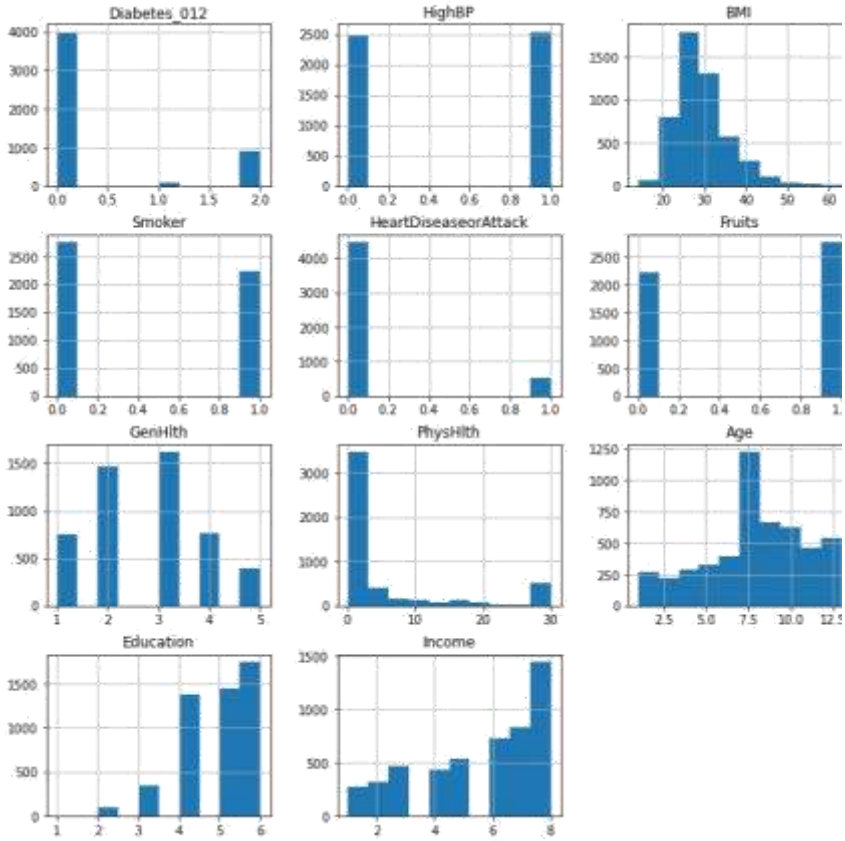
Therefore, after constructing the above-reinforced tree using the XGBoots classifier, we obtain the relevant points for each attribute. Overall, the relevance provides a result that shows how useful or valuable each was each feature in the construction of decision trees grown within the model. The more an attribute is used to make key decisions with decision trees, the greater its relative importance.

Figure 1 Feature importance graph



These are the best-suited columns for prediction as per forwarding Feature Selection. Also, the distribution of the variables taken into account for the second forecast generation is shown below in Figure 2:

Figure 2 Pre-processed data visualization

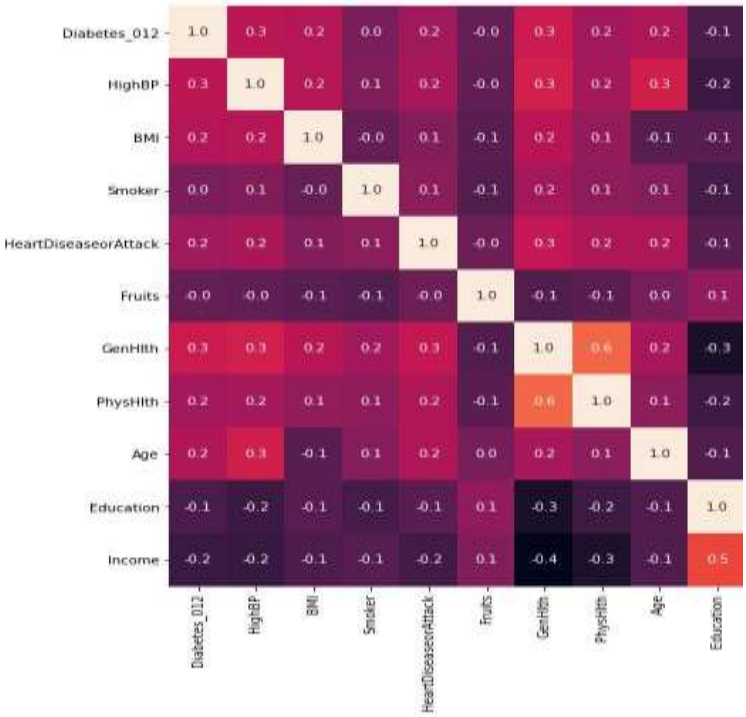


Referring to the histogram graph and the distribution of the remaining variables, in the second step we perform a correlation test between the remaining variables. Method fails because the condition was to select those variables that cannot be correlated with each other. For each of the methods, we have then calculated the accuracy of the model in Table 1. Furthermore, from the combination of the GA method with the GSA method, we obtain a hybrid method.

Table 1. Comparison of classification models

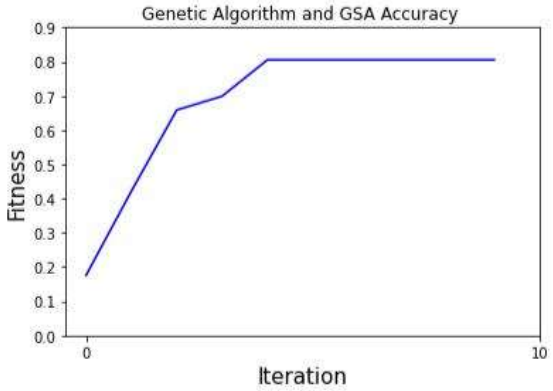
Algorithm	Accuracy
Filter Methods	70%
Wrapper methods	77%
Genetic Algorithm	78%
GA-GSA Hybrid	80%

Since the genetic algorithm has the highest accuracy value, in combination with the GSA method we have obtained an increase in accuracy of 2%.



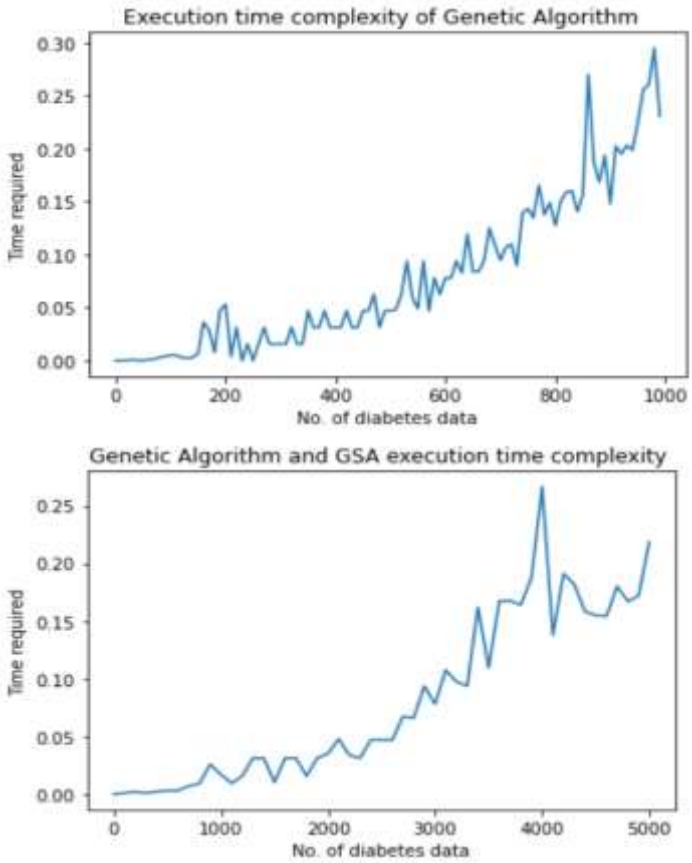
Once we have obtained also the result of the accuracy of the model, what we are interested in next is how the complexity of the algorithm changes. We, therefore, have tested the complexity of the genetic algorithm and then the execution time of the hybrid method created by GA and GSA.

Figure 3 Accuracy for the hybrid model



Here we see that there is a close correlation between some of the variables, which shows me that the filter

Figure 5 GA vs Hybrid GA and GSA



We observe that the complexity is O (n^2) and when we use the hybrid method, the complexity we computed is O(nlog(n)). If the dataset value increases regularly, we have saved time.

Conclusion

The identification and processing of data for chronic diseases such as diabetes will bring information and greatly influence the timely treatment of the disease. Machine Learning methods were used in the data processing process giving a prediction of the patient's illness and state of health. Being a disease with the highest mortality prediction is a challenging process for medicine to provide the solution in real-time and the use of technology would bring a reduction in the number of fatalities. The comparison between the three metaheuristic algorithms and finding the highest accuracy value helps us to use an even safer method. As we said later we would use the intertwined GSA method. The proposed hybrid approach between the two ranking methods as GA and GSA brought a high level of security with almost 80% accuracy and at the same time significantly improved the execution time of the algorithm. If we are going to make a summary and comparison

between other optimization methods by referring to the table below:

Table 2. Comparison between the accuracy of algorithms

Non metaheuristic algorithm	Accuracy	Metaheuristic algorithm	Accuracy
SVM	98.9%	Filter method	70%
kNN	91%	Wrapper method	77%
Random Forest	83%	Genetic algorithm	78%
NaivBayes	89.3%	Hybride GA and GSA	80%
Decision Tree	89%		
Logistic regresion	98%		

The accuracy of the optimization algorithms is calculated as an average approximately of 150 items considered in the study from 2015 to 2021. While the accuracy of meta-heuristic algorithms is obtained from this study. Compared to one another other, meta-heuristic algorithms do not display the same level of accuracy, but their advantage is that as the number of data elements increases, they do not degrade. They can be used for any kind of issue, usually have an exponential runtime, use data from the Real-world, etc. [22] Later on this article will advance in finding other methods that provide even greater and more efficient accuracy. Alongside the combination of hybrid methods, will be used new proposals for setting the coefficients of the distances considered in the analysis.

Bibliography

1. I. Newaz, N. I. Haque, A. K. Sikder, M. A. Rahman and A. S. Uluagac, " Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems," GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020, pp. 1-6, doi:10.1109/GLOBECOM42002.2020.9322472.
2. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, " Secure and Robust Machine Learning for Healthcare: A Survey," in IEEE Reviews in Biomedical Engineering, vol. 14, pp. 156-180, 2021, doi: 10.1109/RBME. 2020.3013489.
3. Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, Marzyeh Ghassemi, Ethical Machine Learning in Healthcare, Journal Article, 2021, Annual Review of Biomedical Data Science, 10.1146/annual-biodatasci-092820-114757 [doi], https://www.annualreviews.org/doi/abs/ 10.1146/ann rev-biodatasci-092820-114757

4. Zeeshan Ahmed, Khalid Mohamed, Saman Zeeshan, XinQi Dong, Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine, Database, Volume 2020, 2020, baaa010, https://doi.org/10.1093/database/baaa010
5. Maniruzzaman, M., Rahman, M.J., Ahammed, B. *et al.* Classification and prediction of diabetes disease using machine learning paradigm. *HealthInfSciSyst* **8,** 7 (2020). https://doi.org/10.1007/s13755-019-0095-z
6. P. Mohamed Jebran and S. Gupta, "Microaneurysm detection by multiple feature subset extraction and selection based on SVM-weights and Genetic Algorithm-Neural Network," *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2021, pp. 129-134, doi: 10.1109/ICACCS51430.2021.9441746.
7. V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus — A machine learning approach," *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 2015, pp. 122-127, doi: 10.1109/RAICS. 2015.7488400.
8. Hoque, N., Singh, M. & Bhattacharyya, D.K. EFS-MI: an ensemble feature selection method for classification. *Complex Intell. Syst.* **4,** 105– 118 (2018). https://doi.org/10.1007/s40747-017-0060-x
9. Aishwarya Mujumdar, V Vaidehi, Diabetes Prediction using Machine Learning Algorithms, Procedia Computer Science, Volume 165, 2019, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.047.
10. Maniruzzaman, M., Rahman, M.J., Al-MehediHasan, M. *et al.* Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. *JMedSyst* **42,** 92 (2018). https://doi.org/10.1007/s10916-018-0940-7
11. Battineni, G.; Sagaro, G.G.; Nalini, C.; Amenta, F.; Tayebati, S.K. Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. *Machines* **2019**, *7*, 74. https://doi.org/10.3390/machines7040074
12. Mahsa Alirezaei, Seyed Taghi Akhavan Niaki, Seyed Armin Akhavan Niaki, A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines, Expert Systems with Applications, Volume 127, 2019, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.02.037.
13. Reza, M. *et al.* (2021). Automatic Diabetes and Liver Disease Diagnosis and Prediction Through SVM and *K*NN Algorithms. In:

Hassanien, A.E., Bhattacharyya, S., Chakrabati, S., Bhattacharya, A., Dutta, S. (eds) Emerging Technologies in Data Mining and Information Security. Advances in Intelligent Systems and Computing, vol 1300. Springer, Singapore. https://doi.org/10.1007/978-981-33-4367-2_56

14. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019. 8965556.
15. Wang, X., Zhai, M., Ren, Z. *et al.* Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC Med Inform Decis Mak* **21,** 105 (2021). https://doi.org/10.1186/s12911-021-01471-4
16. Varma, Kucharlapati Manoj, and Dr BS Panda. "Comparative analysis of Predicting Diabetes Using Machine Learning Techniques." *J. Emerg. Technol. Innov. Res* 6 (2019): 522-530.
17. https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c,
18. Ahmad Alsahaf, Nicolai Petkov, Vikram Shenoy, George Azzopardi, A framework for feature selection through boosting, Expert Systems with Applications, Volume 187, 2022, 115895, ISSN 0957-4174,https://doi.org/ 10.1016/j.eswa.2021.115895.
19. Ibrahim, Nuhu & Hamid, H.A. & Rahman, Shuzlina & Fong, Simon. (2018). Feature selection methods: Case of filter and wrapper approaches for maximizing classification accuracy. Pertanika Journal of Science and Technology. 26. 329-340.
20. oo, J., Abdullah, A.R. A new and fast rival genetic algorithm for feature selection. *JSupercomput* **77,** 2844–2874 (2021). https://doi.org/10.1007/ s11227-020-03378-9
21. Y. Xiang, S. Gubian, and F. Martin, "Generalized Simulated Annealing", in Computational Optimization in Engineering - Paradigms and Applications. London, United Kingdom: IntechOpen, 2017 [Online]. Available: https://www.intechopen.com/chapters/52820. doi: 10.5772/66071
22. Harvinder Singh, Sanjay Tyagi, Pardeep Kumar, Sukhpal Singh Gill, Rajkumar Buyya, Metaheuristics for scheduling of heterogeneous tasks in cloud computing environments: Analysis, performance evaluation, and future directions, Simulation Modelling Practice and Theory, Volume 111, 2021, 102353, ISSN 1569-190X, https://doi.org/10.1016/j.simpat.2021. 102353.
23. Ye, F. Evolving the SVM model based on a hybrid method using swarm optimization techniques in combination with a genetic algorithm for medical diagnosis. *Multimed Tools Appl* **77,**

3889–3918 (2018). https://doi.
org/10.1007/s11042-016-4233-1

24. Role of machine learning algorithms over heart diseases prediction, AIP Conference Proceedings 2292, 040013 (2020); https://doi.org/10.
1063/5.0030743,SivaKumar Jonnavithula*,* Abhilash Kumar Jha*,* Modepalli Kavitha*, and* Singaraja