Machine Learning based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach using Correlation Matrix with Heatmap and SFS

Selim Buyrukoglu and Ayhan Akbas

***Abstract*—A new hybrid machine learning method for the prediction of type 2 diabetes is introduced and explained in detail. Also outcomes are compared with the similar researches. Early prediction of diabetes is crucial to take necessary measures (i.e. changing eating habits, patient weight control etc.), to defer the emergence of diabetes and to reduce the death rate to some extent and ease medical care professionals' decision making in preventing and managing diabetes mellitus.The purpose of this study is the creation of a new hybrid feature selection approach combination of Correlation Matrix with Heatmap and Sequential forward selection (SFS) to reveal the most effective features in the detection of diabetes. A diabetes data set with 520 instances and seven features were studied with the application of the proposed hybrid feature selection approach. The evaluation of the selected optimal features was measured by applying Support Vector Machines(SVM), Random Forest(RF), and Artificial Neural Networks(ANN) classifiers. Five evaluation metrics, namely, Accuracy, F-measure, Precision, Recall, and AUC showed the best performance with ANN (99.1%), F-measure (99.1%), Precision (99.3%), Recall (99.1%), and AUC (99.8%). Our proposed hybrid feature selection model provided a more promising performance with ANN compared to other machine learning algorithms.**

***Index Terms*—Artificial Neural Network, Correlation Matrix, Sequential Forward Selection, Diabetes Mellitus, Hybrid Feature Selection**

I. Introduction

DIABETES Mellitus (DM), commonly known as diabetes, is a metabolic disease that causes higher-than-normal blood sugar level. The insulin hormone secreted by the body carries excess sugar in the blood to your cells to be stored or used for energy. In diabetes, the body either does not produce enough insulin or cannot use insulin effectively to reduce the glucose level in the blood. As a result, uncontrolled high blood sugar can damage the nerve, eyes, kidneys, and other organs [1]. The general symptoms of diabetes include increased hunger, increased thirst, weight loss, frequent urination, blurry vision, extreme fatigue and sores that don't heal. Pre-diabetes occurs when your blood sugar is higher than normal, but usually it's not high enough for a diagnosis of type 2 diabetes. Therefore, early prediction of diabetes is crucial to take necessary measures (i.e. changing eating habit, weight control etc.), to defer the emergence of diabetes and to reduce the death rate to some extent.

Machine Learning (ML), a domain of artificial intelligence (AI), makes use of capabilities of computers to learn automatically without explicitly programmed and improve from experience. Machine learning in healthcare is becoming more widely used and is helping clinicians in many different ways [2]. One of the most common use cases in medicine for machine learning is the diagnosis decision support [3].The power of machine learning models is its ability to process huge datasets beyond the scope of human capability [4], and then reliably convert analysis of that data into clinical insights to provide better information to doctors at the point of decision making. There are plenty of studies on Diabetes that employed a machine learning (ML) based models in the literature as systematically reviewed by De Silva et. al. (2020) for various databases[5]. In our research, we developed a hybrid model for selecting most dominating features so as to maximize the output accuracy. The data analyzed in this paper, has been collected from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh (2020) [6]. Features in this dataset are listed and described in Table I.

In the literature, there are numerous studies on the diagnosis of diabetes mellitus. With advances in computer sciences, machine learning techniques have become common in healthcare. Overview of studies employing ML models in their researches has been published by Chaki et. al. (2020) [7] and Kavakiotis et. al.(2017) [8] who conducted systematic reviews on the studies on machine learning and artificial intelligence based Diabetes Mellitus detection and provided a detailed overview of DM detection. Another group of researches, Jashwanth et. al.(2020) [9] studied the performance of six machine learning classifiers (namely, support vector

SELİM BUYRUKOGLU is with the Department of Computer Engineering, Engineering Faculty, Cankiri Karatekin University, Cankiri, Turkey e-mail: (sbuyrukoglu@karatekin.edu.tr)
https://orcid.org/0000-0001-7844-3168
AYHAN AKBAS is with the Department of Computer Engineering, Engineering Faculty, Abdullah Gul University, Kayseri, Turkey e-mail: (ayhan.akbas@agu.edu.tr)
https://orcid.org/0000-0002-6425-104X

Manuscript received July 19, 2021; accepted January 4, 2022
DOI: : 10.17694/bajece.973129

machine, K-nearest neighbors, logistic regression, naive bayes, gradient boosting and random forest) on Pima Indian Diabetes Database which is a common dataset for diabetes studies. The performance of all the six classifiers were compared using Accuracy score, Receiver Operating Curve (ROC), Precision, Recall, F-measure evaluated from each model. Random Forest classifier has shown the highest performance compared to remaining classifiers used in their proposed methodology. Lai et. al [10] also developed a predictive model to predict the risk for developing DM using Logistic Regression and Gradient Boosting Machine (GBM) techniques in Canadian patients. Fasting blood glucose, body mass index, high-density lipoprotein, and triglycerides were the most important predictors in their model. Accuracy for the proposed GBM model is 84.7% and for the proposed Logistic Regression model is 84.0%. In another study [11], classifiers (Decision Tree (DT), Artificial Neural Networks (ANN), Logistic Regression (LR) and Naive Bayes (NB)) were compared for the risk of diabetes prediction, then bagging and boosting techniques were investigated for improving the robustness. Random Forest (RF) algorithm was suggested as the best performance of disease risk classification. To the best of our knowledge, our approach is novel in terms of the selection of features and employment of hybrid combination of classifier to achieve the most accurate outcome. Most of the studies in the literature utilized single classifier models contrary to our approach. Moreover, we used a new dataset consisting of all diabetes symptoms unlike the great majority of studies on early prediction of diabetes that employed pima indian diabetes dataset[12] where most related symptoms are missing and features in the dataset are very limited.

Other research using deep learning techniques has been carried out by Swapna et al. [13]. They utilized the RR-interval signals known as heart rate variability (HRV) signals (derived from electrocardiogram (ECG) signals) and used for the non-invasive detection of diabetes. In their research, they employed long short-term memory (LSTM), convolutional neural network (CNN) and their combinations for extracting complex temporal dynamic features of the input HRV data which was then, passed to SVM (Support Vector Machine) for classification. The system proposed is able to diagnose diabetes using ECG signals with an accuracy of 95.7%. A similar study that emplyed machine learning approach was carried out by Zou et. al. [14]. In their study, they used decision tree, random forest and neural network to predict diabetes mellitus in people. The dataset they used is obtained from a hospital physical examinations and contains 14 attributes. Principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) are employed to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used.

In the paper, we introduced the area of the study, mentioning the similar researches in section I. In the section II, the proposed approach is given together with the data collection and preprocessing steps. Here, classifiers used in the algorithms are summarized briefly. Experimental results and implementation are given and discussed in section III. Finally, section IV includes conclusions and future directions of the study.

II. Proposed Approach

In this study, we have implemented a new hybrid feature selection approach. Fig. 1 illustrates the proposed architecture which consists of several phases of data processing steps.

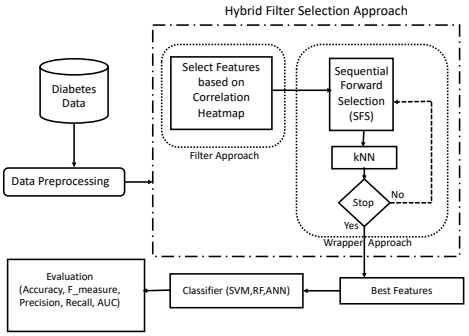


Fig. 1: Architecture of the proposed approach

A. Data Collection and Preprocessing

The diabetes data set is collected from patients in Sylhet Diabetes Hospital in Sylhet, Bangladesh [6]. It has 520 instances (342 confirmed Diabetes positive and 178 Diabetes negative) and also contains seventeen features and two classes. Description of datasets is presented in Table I. The type of these features are categorical except only age feature (numeric).

The aim of data collection here is to predict diabetes disease applying a new hybrid feature selection approach. The importance of the using features in the prediction of diabetes disease was already specified in Section I. The most important features in the prediction of diabetes disease will be highlighted through the proposed hybrid feature selection approach. There is on missing value and all observations are complete.

B. Correlation Matrix with Heatmap

Features relation to each other or the target feature is provided through correlation matrix as illustrated in Fig. 2. Correlation matrix is a graphical representation of data to identify which features are most related to the target feature. Each feature in a dataset is represented as colors which mean that the colors notice information to researchers in terms of the correlation between features. We created our heatmap of a correlation matrix which is presented in 5 steps in Algorithm 1.

C. Sequential Forward Selection (SFS)

Feature selection algorithms are used to eliminate unusable and redundant features. [15] highlighted that feature selection is an efficient way in terms of dimensional reduction of data with high dimensionality. Sequential forward selection (SFS) algorithm was applied to select an optimal subset of data from the 17 features. The reason behind the selection of SFS in this study is that wrapper feature selection approaches are widely

Feature	Description	Value	Ratio
Age	Age of patient	Numeric	Range:16-90 Mean:48, Std.Dev:12.15
Gender	Sex	Male/Female	63.1% / 36.9%
Polyuria	Frequent urination	Yes/No	50.4% / 49.6%
Polydipsia	Excessive thirst	Yes/No	55.2% / 44.8%
Sudden Weight Loss	Uncontrolled weight loss	Yes/No	58.3% / 41.7%
Weakness	Condition of being weak	Yes/No	58.7% / 41.3%
Polyphagia	Abnormal desire to consume excessive amounts of food	Yes/No	54.4% / 45.6%
Genital thrush	Genital fungal infections	Yes/No	77.7% / 22.3%
Visual Blurring	Unclear vision	Yes/No	55.2% / 44.8%
Itching	Irritating sensation of the skin	Yes/No	51.3% / 48.7%
Irritability	Become easily anger	Yes/No	75.8% / 24.2%
Delayed healing	Impaired healing	Yes/No	54.0% / 46.0%
Partial paresis	Slight/Partial Paralysis	Yes/No	56.9% / 43.1%
Muscle stiffness	Status that muscles feel tight and more difficult to move	Yes/No	62.5% / 37.5%
Alopecia	Hair Loss	Yes/No	65.6% / 34.4%
Obesity	Obesity	Yes/No	83.1% /16.9%
Class (Target)	Diabetes Status	Positive/Negative	61.5% / 38.5%

TABLE I: Description of Features in the dataset

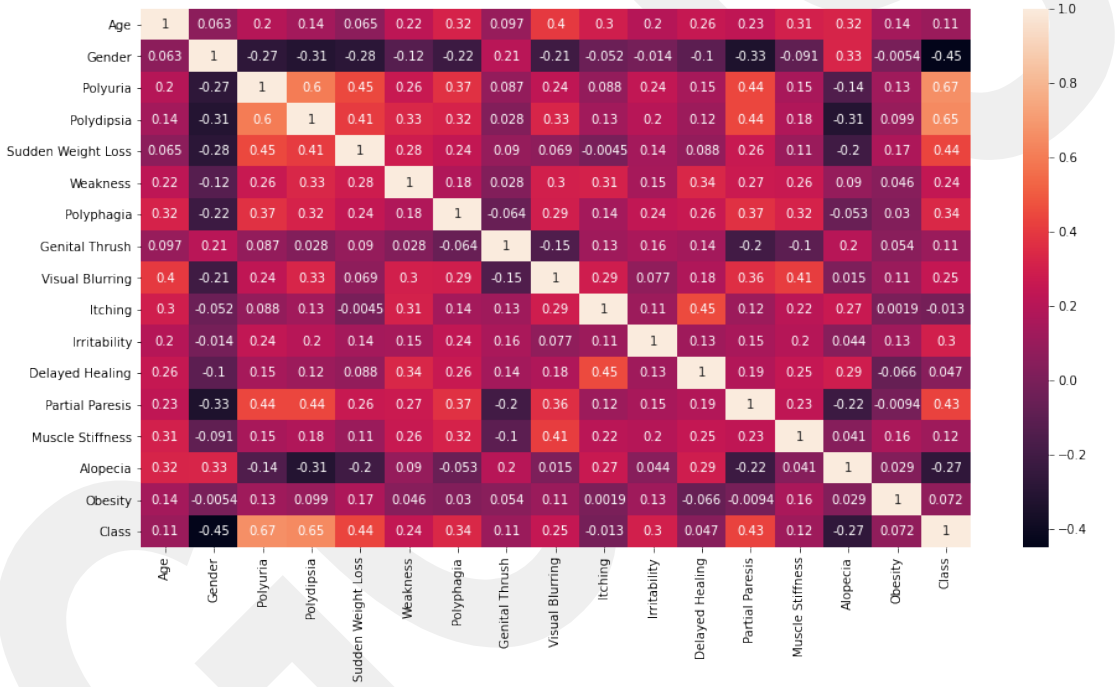


Fig. 2: Heatmap of the Correlation Matrix for the Diabetes Data

Algorithm 1: Steps of Creation of Correlation Matrix with Heatmap

1. Import Data;
2. Identify Independent Columns (X) and Target Column (Y);
3. Create Correlation Matrix (get correlations of each features in dataset);
4. Create Heatmap with size (17, 17);
5. Plot and Export Heatmap;

used to detect fundamental interaction between features and also SFS is a user-friendly wrapper approach. An empty set is created by the SFS algorithm initially. The feature has the highest feature importance score is added to the empty set. The features are ranked based on feature importance to extract the top features using extra tree classifier. Then, the feature has the second-highest score is added to the set. The feature adding process to the set continues until the classifier can no longer improve the performance. The k-nearest neighbors (KNN) classifier is implemented in this study since it is one of the most effective machine learning algorithm in terms of classification [16]. Once the best classification performance is achieved, the added features to the set are utilized for classification since they work best together. Algorithm 2 presents pseudo-code for the SFS algorithm.

Algorithm 2: Sequential Forward Selection Pseudo Code

Input: $Y = y_1, y_2, ..., y_d$ whole d-dimensional feature set are taken as input

Output: $X_k = x_j \mid j = 1, 2, ..., k; \quad x_j \ \epsilon \ Y$, where $k = (0, 1, 2, ..., d)$

SFS returns a subset of features; the number of selected features k, where $k < d$, has to be specified a priority.

1. Create an empty set: $X_0 = \phi$, $k = 0$
   We initialize the algorithm with an empty set ϕ ("null set") so that $k = 0$ (where k is the size of the subset).
2. Select best remaining feature:
   x^+= arg max $J(X_k + x)$, where $x \phi Y - X_k$
   $X_{k+1} = X_k + x^+$
   $k = k + 1$
   Goto Step 2
3. Termination: $k = p$

D. Classifiers

Support Vector Machines (SVM): Support Vector Machine (SVM) is an associated machine learning algorithm mostly used for classification and regression analysis [17]. Given a set of training examples, An SVM maps training examples to points in space so as to maximize the width of the gap between the two categories. In other words, a support vector machine can be defined as a vector space-based machine learning method that finds a decision boundary between two classes that are the furthest from any point in the training data. Support vector machines are mostly used to separate binary classification data, for example separating each data in a data set into female or male.

Random Forest (RF): Random forests (RF) or random decision forests is a collective learning method that estimates class (classification) or number (regression) according to the type of the problem by creating a large number of decision trees during the training phase for classification, regression and other tasks [18]. Random decision forests address the problem of decision trees over fitting training sets. Different number of trees (50, 100, 250, 500) was tested in the prediction of type 2 diabetes. The best performance was obtained with five hundred trees.

Artificial Neural Network (ANN): Artificial neural networks (ANN) is a computing technology inspired by the information processing technique of the human brain[19]. With ANN, the way the simple biological nervous system works is imitated. That is to say, it is digital modeling of biological neuron cells and the synaptic bond that these cells establish with each other. Neurons connect to each other in various ways to form networks. These networks have the capacity to learn, memorize and reveal the relationship between data. In other words, ANNs provide solutions to problems that normally require a person's natural abilities to think and observe. The main reason why a person can produce solutions to problems that require thinking and observing skills is the ability of the human brain, and therefore the human being, to learn by living or experimenting. In this study, one input layer with 7 nodes and one hidden layer with 14 neurons were tested.

E. Evaluation Metrics

The evaluation of the machine learning algorithm is always a vital part of every project. When the model is evaluated using one criterion, it may yield satisfactory results, but poor results may be obtained when compared with other criteria or metric. In our study, We mostly use classification accuracy to measure the performance of a model, but this is not enough to reach a solid conclusion. Mostly used classifiers are given here.

Accuracy: It is the ratio of number of correct predictions to the total number of input samples. Accuracy is a good measure when the target classes in the data are nearly balanced.

F-Measure: F-measure or F1 Score, also known as the Sørensen–Dice coefficient or Dice similarity coefficient(DSC), is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

10-fold Cross validation: Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

Precision andn Recall: Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance [20].

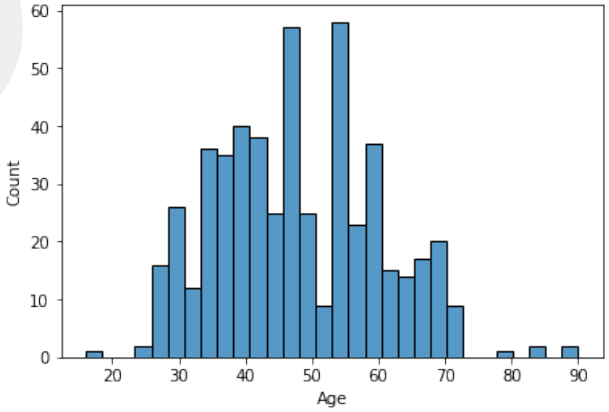


Fig. 3: Patient Age Distribution

III. Analysis and Discussion

Filter and wrapper feature selection methods were embedded to increase the efficiency and accuracy of the classification models. Diabetes data (see Section 2.1) were used for the purposes of this study.

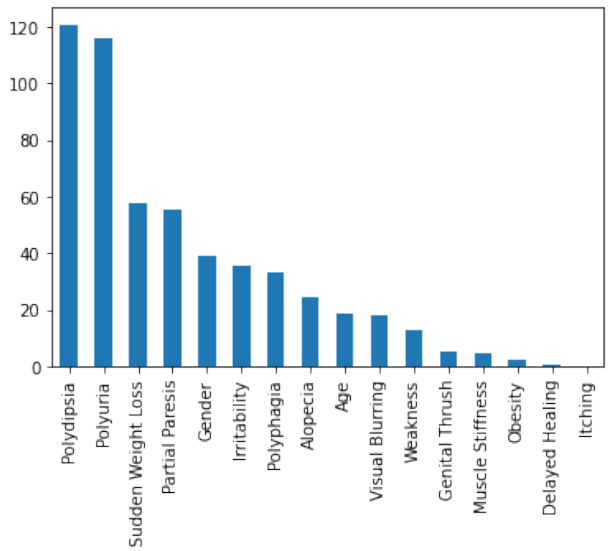


Fig. 4: Feature Score with respect to their contribution towards Class

A. Implementation

Two phases (filter and wrapper) were proposed in the proposed hybrid feature selection approach. In the first phase, correlation matrix with heatmap filter method was used to rank features between the features and target class. A threshold value of correlation heatmap scores was determined to select k-best features. Ideally, features are selected if the correlation score between two features and target class is more than 0.5. However, there are only the features Polyuria and Polydipsia have more than 0.5 correlation score with the target class. Also, most of the features have correlation scores between ˜0.1 and 0.5. Thus, the threshold value of correlation heatmap scores was determined as ˜0.1. Then, 11 features were selected at the end of the first phase and these features are used in the second phase as input. A heatmap of the correlation matrix for the diabetes data is presented in Fig. 2. The correlation between features and target class higher than 0.5 are displayed in dark green which are Polyuria and Polydipsia. These two features play an important role in the detection of diabetes and that the two features show correlation is therefore not surprising.

In the second phase, sequential forward selection (SFS) approach was implemented for selecting the k-best features. The optimal number of features was selected based on feature importance method. It ranks features and gives scores for each feature of the data. Future importance method is a three-based classifier extracting the k-best number of feature. Finally, SFS wrapper feature selection method was applied to select the optimal feature set. At the end, 7 features were selected over 11 features. The evaluation of the selected optimal features were measured by applying Support Vector Machine, Random Forest and Artificial Neural Network classifiers. Five evaluation matrices were used in the evaluation of the proposed approach such as Accuracy, F-measure, Precision, Recall and AUC. Table II presents the classification results for the proposed hybrid feature section approach (both CM + SFS), original feature set, correlation matrix with heatmap (only correlation matrix with heatmap feature selection was applied), SFS (only SFS was applied)

The results obtained from the data in Table II that ANN has the best accuracy rate (99.1%) using 7 features selected based on the proposed hybrid approach while SVM has the lowest accuracy rate (95.8%). Overall, ANN has the best accuracy rate based on not only hybrid method but also Sequential Forward Selection (SFS), Confusion Matrix (CM) and Original (96.5%, 94.6%, 95.6% respectively). As highlighted in Section III, 10-fold cross-validation was used to test the performance of the classification algorithms. As can be seen from Table II that 7 features were selected based on the proposed hybrid approach. These features are set out in Table III.

B. Results

Fig. 5 illustrates the AUC results for SVM, RF and ANN classifiers based on different feature sets. From Fig. 5, it is observed that ANN has the highest AUC rates for each feature set (purple color). Also, The AUC rates of ANN for hybrid method (99.8%) when compared with original features (98.8%), correlation heatmap (99.0%) and SFS (99.6%) methods. Additionally, SVM and RF classifiers achieved to obtain highest AUC rates with hybrid method (98.5%, 99.7% respectively).

C. Comparison of previous studies and the proposed model

Table IV compares the results obtained from the previous studies and the proposed model for prediction of diabetes. The purpose of the studies is the same with our study. Anshuman Guha [21], Ahmed Kareem et al [22], Kezban Alpan [23] and Jingyu Xue [24] used the same data set with the data set used in this study consisting of 520 instances. Ahmed Kareem et al. achieved to obtain 98.8% accuracy applying RF value without applying any feature selection technique. Kezban Alpan (2020) also accomplished 98% accuracy applying kNN with nine selected features. These features were selected based on Gain Ratio feature selection technique. However, there is no deatiled information about how these models were fitted in these studies. On the other hand, various diabetes data were used in different studies and there are four studies achieved to reach more than 90% accuracy. For instance, Tapak et al.[25] applied the RF using 2000 diabetes samples and achieved 98.6% accuracy. Also, the study of D. Jashwanth Reddy [26] presented the successful result applying SVM using 6500 samples (98.6%). Furthermore, the studies of Abbas et. al. 2019 and Aishwarya Mujumdar [27] provided higher accuracy value (96.8% and 96% respectively).

IV. Conclusion and Future Directions

Early detection of diabetes plays an importnat role in treatment. Datasets consist of many features and they can be chosen using significant feature selection methods to build a

Classifier	Method	Selected Features	Accuracy	F-measure	Precision	Recall
SVM	Original	16	0.903	0.823	0.823	0.823
	Heatmap	11	0.911	0.856	0.857	0.865
	SFS	9	0.948	0.885	0.886	0.890
	Hybrid	7	0.958	0.921	0.922	0.924
RF	Original	16	0.906	0.906	0.906	0.906
	Heatmap	11	0.925	0.926	0.928	0.925
	SFS	9	0.963	0.963	0.964	0.963
	Hybrid	7	0.988	0.988	0.988	0.988
ANN	Original	16	0.956	0.956	0.957	0.956
	Heatmap	11	0.946	0.946	0.947	0.946
	SFS	9	0.965	0.965	0.965	0.965
	Hybrid	7	0.991	0.991	0.993	0.991

TABLE II: Comparison of Performances of Classifiers

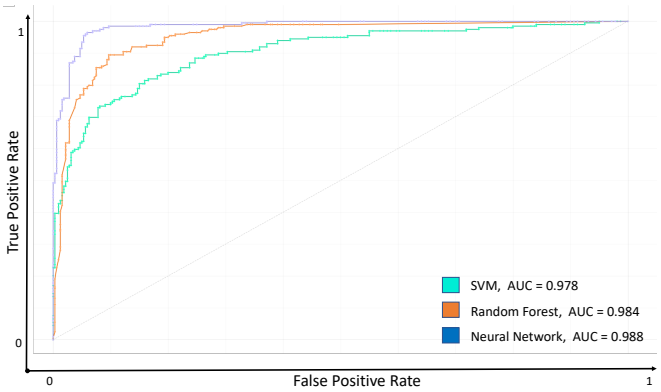
Feature ID	Feature
4	Polydipsia
3	Polyuria
5	Sudden Weight Loss
13	Partial Paresis
11	Irritability
7	Polyphagia
1	Age

TABLE III: Selected features with hybrid approach

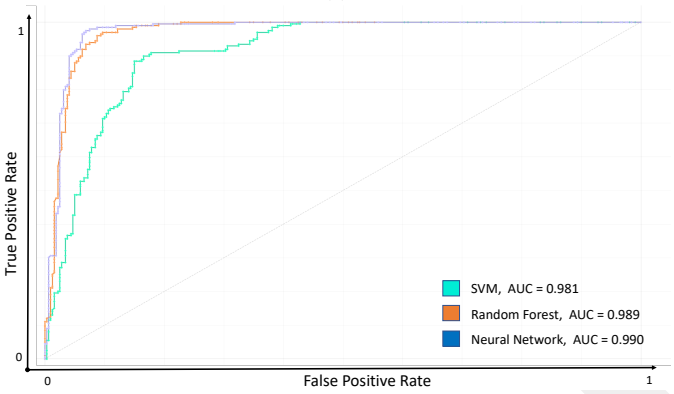
Studies	Data Size	Classifier	Accuracy
Anshuman Guha [21]	520	RF	94.8%
Ahmed Kareem et al. [22]	520	RBF	98.8%
Kezban Alpan [23]	520	kNN	98.0%
Jingyu Xue [24]	520	SVM	96.5%
D. Jashwanth Reddy [26]	2000	RF	98.48%
Maniruzzaman et al.[28]	768	LDA,QDA,NB,GPC	81.9%
Deng and Kasabo [29]	768	ESOM	78.4%
Christobel and Sivaprakasam [16]	768	KNN	78.1%
Farahmandian et al. [30]	768	SVM,KNN,NB,ID3,CART,c5.0	81.0%
Khashei et al.[31]	68	LDA,QDA,KNN,SVM,ANN,HPM	80.0%
Nongyao Nai-arun [32]	30112	RF	85.5%
Aishwarya Mujumdar [27]	800	Logistic Regression	96.0%
Tapak et al. [25]	6500	SVM	98.6%
Abbas et al. [33]	1492	SVM-RBF	96.8%
Proposed Approach I	520	Hybrid + ANN	99.1%
Proposed Approach II	520	Hybrid + RF	98.8%

TABLE IV: Results from Previous Studies

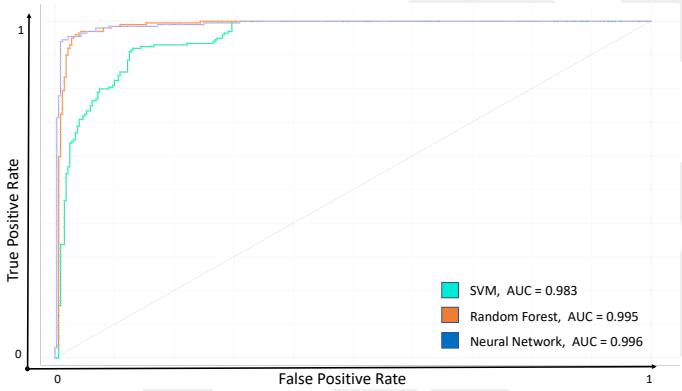
well predictive model. The reason for this is that all feature selection methods cannot take into account important features to enhance the predicting process. Note that significant features are used to build well predictive models that help health-care professionals to treat patients. A reasonable approach to tackle this issue could be to describe different feature selection approaches. In this case, a predictive model has been built based on a new hybrid feature selection approach. The proposed hybrid feature selection approach combines filter (correlation matrix with heatmap) and wrapper (sequential forward selection) methods. Initially, 11 features were selected applying correlation matrix with heatmap over 16 features. Then, sequential forward selection (SFS) approach was implemented for selecting the k-best features using these 11 features, and so 7 features were selected through SFS. Three different machine learning algorithms applied to the selected features (feature subsets) and then their performances were compared to reveal the efficiency of the proposed new hybrid feature selection approach on the diabetes detection. This study revealed that the best classification accuracy (99.1%) is obtained by applying the Artificial Neural Network algorithm with feature set generating from the proposed hybrid feature selection approach. Finally, we can see that machine learning algorithms and the proposed hybrid feature selection approach have made outstanding contributions in the diabetes data. Our future research will be about the combination of feature selection approach and deep learning algorithms to improve classification accuracy based on the image dataset.



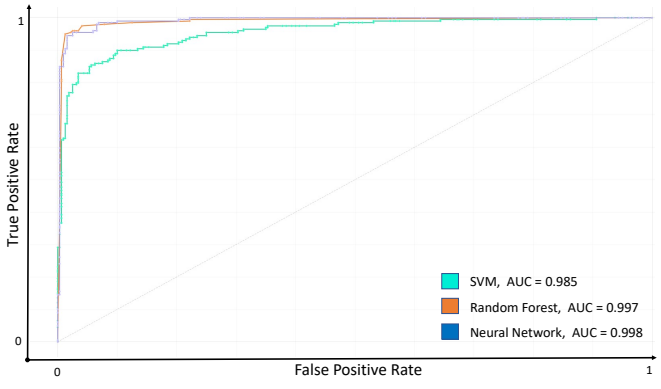
(a)



(b)



(c)



(d)

Fig. 5: ROC curve of (a) Original, (b) Heatmap, (c) SFS and (d) Hybrid

REFERENCES

[1] Stephanie Watson, "Everything You Need to Know About Diabetes," 2020. [Online]. Available: https://www.healthline.com/health/diabetes

[2] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, "Machine learning in healthcare: A review," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 910–914.

[3] N. Peiffer-Smadja, T. Rawson, R. Ahmad, A. Buchard, G. Pantelis, F.-X. Lescure, G. Birgand, and A. Holmes, "Machine learning for clinical decision support in infectious diseases: A narrative review of current applications," *Clinical Microbiology and Infection*, vol. 26, 09 2019.

[4] E. Sevinc, "A novel evolutionary algorithm for data classification problem with extreme learning machines," *IEEE Access*, vol. 7, pp. 122 419–122 427, 2019.

[5] K. D. Silva, W. K. Lee, A. Forbes, R. T. Demmer, C. Barton, and J. Enticott, "Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis," *International Journal of Medical Informatics*, vol. 143, no. August, p. 104268, 2020. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2020.104268

[6] H. Zheng, H. W. Park, D. Li, K. H. Park, K. H. Ryu, J. Xue, F. Min, F. Ma, N. P. Tigga, S. Garg, Stephanie Watson, I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, D. Jashwanth Reddy, B. Mounika, S. Sindhu, T. Pranayteja Reddy, N. Sagar Reddy, G. Jyothsna Sri, K. Swaraja, K. Meenakshi, P. Kora, M. M. F. Islam, R. Ferdousi, S. Rahman, H. Y. Bushra, S. Gupta, A. Guha, D. Jain, V. Singh, A. K. Farahat, A. Ghodsi, M. S. Kamel, J. Chaki, S. Thillai Ganesh, S. K. Cidham, S. Ananda Theertan, V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, K. Alpan, G. S. Ilgi, , K. Akyol, and B. Şen, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," *Knowledge and Information Systems*, vol. 15, no. 3, pp. 113–125, 2020.

[7] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda 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*, 2020. [Online]. Available: https://doi.org/10.1016/j.jksuci.2020.06.013

[8] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017. [Online]. Available: https://doi.org/10.1016/j.csbj.2016.12.005

[9] D. Jashwanth Reddy, B. Mounika, S. Sindhu, T. Pranayteja Reddy, N. Sagar Reddy, G. Jyothsna Sri, K. Swaraja, K. Meenakshi, and P. Kora, "Predictive machine learning model for early detection and analysis of diabetes," *Materials Today: Proceedings*, 2020. [Online]. Available: https://doi.org/10.1016/j.matpr.2020.09.522

[10] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, no. 1, pp. 1–9, 2019.

[11] N. Nai-Arun and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," *Procedia Computer Science*, vol. 69, pp. 132–142, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2015.10.014

[12] Kaggle, "Pima Indians Diabetes Dataset," 2021. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[13] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018. [Online]. Available: https://doi.org/10.1016/j.icte.2018.10.005

[14] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, Nov. 2018. [Online]. Available: https://doi.org/10.3389/fgene.2018.00515

[15] S. Pratama, A. Muda, Y.-H. Choo, and N. Muda, "Computationally inexpensive sequential forward floating selection for acquiring significant features for authorship invarianceness in writer identification," *International Journal of New Computer Architectures and their Applications (IJNCAA)*, vol. 1, pp. 581–598, 01 2011.

[16] Y. A. Christobel and P. Sivaprakasam, "A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset," *International Journal of Engineering and Advanced Technology*, vol. 2, no. 3, pp. 396–400, 2013.

[17] Wikipedia, "Support vector machine," 2021. [Online]. Available: https://en.wikipedia.org/wiki/Support-vector-machine

[18] ——, "Random Forest," 2021. [Online]. Available: https://en.wikipedia.org/wiki/Random_forest

[19] ——, "Artificial Neural Network," 2021. [Online]. Available: https://en.wikipedia.org/wiki/Artificial_neural_network
[20] ——, "Precision and Recall," 2021. [Online]. Available: https://en.wikipedia.org/wiki/Precision_and_recall
[21] A. Guha, "Building Explainable and Interpretable model for Diabetes Risk Prediction," *International Journal of Engineering Research and Technology*, vol. 9, no. 09, pp. 1037–1042, 2020.
[22] A. Kareem, L. Shi, L. Wei, and Y. Tao, "A Comparative Analysis and Risk Prediction of Diabetes at Early Stage using Machine Learning Approach A Comparative Analysis and Risk Prediction of Diabetes at Early Stage using Machine Learning Approach," *International Journal of Future Generation Communication and Networking*, vol. 13, no. 3, pp. 4151–4163, 2020.
[23] K. Alpan and G. S. Ilgi, "Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, oct 2020, pp. 1–7.
[24] J. Xue, F. Min, and F. Ma, "Research on diabetes prediction method based on machine learning," *Journal of Physics: Conference Series*, vol. 1684, no. 1, 2020.
[25] L. Tapak, H. Mahjub, O. Hamidi, and J. Poorolajal, "Real-data comparison of data mining methods in prediction of diabetes in iran," *Healthcare Informatics Research*, vol. 19, no. 3, p. 177, 2013.
[26] D. Reddy, B. Mounika, S. Sindhu, T. Reddy, N. Reddy, G. Sri, K. Swaraja, M. Kollati, and P. Kora, "Predictive machine learning model for early detection and analysis of diabetes," *Materials Today: Proceedings*, 10 2020.
[27] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019. [Online]. Available: https://doi.org/10.1016/j.procs.2020.01.047
[28] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, Jan. 2020.
[29] D. Deng and N. Kasabov, "On-line pattern analysis by evolving self-organizing maps," *Neurocomputing*, vol. 51, pp. 87–103, apr 2003.
[30] M. Farahmandian, Y. Lotfi, and I. Maleki, "Data Mining Algorithms Application in Diabetes Diseases Diagnosis : A Case Study," *MAGNT Research Report*, vol. 3, no. 1, pp. 989–997, 2015.
[31] M. Khashei, S. Eftekhari, and J. Parvizian, "Diagnosing diabetes type ii using a soft intelligent binary classification model," *Review of Bioinformatics and Biometrics*, vol. 1, no. 1, pp. 9–23, 2012.
[32] N. Nai-arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Computer Science*, vol. 69, pp. 132–142, 2015.
[33] H. T. Abbas, L. Alic, M. Erraguntla, J. X. Ji, M. Abdul-Ghani, Q. H. Abbasi, and M. K. Qaraqe, "Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test," *PLOS ONE*, vol. 14, no. 12, p. e0219636, Dec. 2019.

BIOGRAPHIES

Selim Buyrukoglu is an Asst. Prof. in Computer Engineering Department at Cankiri Karatekin University. He completed his PhD at Loughborough University, UK in 2019, He received an MSc degree in Advance Computer Science in 2014 from Leicester University, UK and also a BSc degree in Computer Engineering in 2010 from European University of Lefke, TRNC. He is particularly focusing on the application of Machine Learning and Deep Learning methods on interdisciplinary subjects.

Ayhan Akbas received his B.S. and M.S. degrees from Middle East Technical University in Electrical and Electronics Engineering in 1991 and 1995, Ph.D degree from Computer Engineering from TOBB ETU in 2016, respectively. He is currently working as assistant professor at Computer Engineering Dept. in Abdullah Gul University. His areas of interest are Wireless Sensor Networks, IoT, Wireless Communication, and machine learning.