Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

ENSEMBLE MODEL FOR DIABETIC DIAGNOSIS

RAMBABU PEMULA[1]

Associate professor,
Department of Computer Science Engineering,
Raghu Engineering College,
Visakhapatnam, Andhra Pradesh.
rpemula@gmail.com

P.GOWTHAMI[2] 19981A05D3@raghuenggcollege.in

S.SRAVYA GEETHIKA[3] 19981A05E3@raghuenggcollege.in

P. CHANDRIKA[4] 19981A05D1@raghuenggcollege.in

*[*2,3,4] 4th year Btech Students, Department of Computer Science Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh.*

ABSTRACT:

Diabetes is a complaint in which the body's capability to produce or respond to the hormone insulin is bloodied, performing in abnormal metabolism of carbohydrates and elevated situations of glucose in the blood. Without mortal backing, hidden patterns can be set up using machine literacy. The suggested methodology intends to create an intelligent agent to detect whether a person is having diabetes or not. A dataset including around 769 records consisting of 8 features is used. This model is aimed to construct an ensemble method with a classification algorithm K-Nearest Neighbors (KNN). The presence or absence of diabetes from the characteristics is also determined using the K Nearest Neighbor (KNN).

KEY WORDS:
Diabetes, Ensemble Model, Classification, K-Nearest Neighbor Classifier

1. INTRODUCTION:

The World Health Organization (WHO) reports that there are more than 400 million people who are suffering from diabetes, and it becomes the cause of death in nearly 1.6 million [1]. According to estimates, breast cancer caused the deaths of approximately 214,360 women in China by the year 2008, and this number is projected to increase to nearly 3 million by starting of 2022 [2]. Because of the situation, there are numerous families suffer along with the patients [3]. Therefore, identifying the reasons behind "such a large amount of deaths" is crucial. According to world reports it is noticed that early detection of disease helps patients to live longer life [4]. In general, ML techniques can effectively separate the useful information from various sources [5-8]. Moreover, ML can be used to streamline diagnostic procedures, allowing healthcare professionals to make more accurate diagnoses in a shorter amount of time. This can lead to more efficient and effective treatment plans for patients, improving overall healthcare outcomes. Various classification methods have been utilized in the medical field such as K-Nearest Neighbors, and Artificial Neural Networks and Most of them show tremendous achievement in performance. These models mainly focused on accuracy, while neglecting the balancing of the data. Whenever faced with imbalanced data, the

classifier may give more weight to the majority class and may struggle to accurately classify minority class samples. This can result in a decision boundary that is biased toward the minority class, as the classifier tries to account for the underrepresented samples. Therefore, it can become troublesome in the performance of a classifier, mainly in the medical diagnosis application. Inspired by the aforementioned limitation, our research focuses solely on investigating the binary classification task and presents an ensemble approach to diagnose diabetes in the presence of imbalanced data. The proposed method involves three distinct phases.

In the preliminary stage, we will be introducing a method called Synthetic Minority Oversampling Technique (SMOTE) to resample the data points. SMOTE is used to balance the dataset. This is because SMOTE generates new synthetic instances for the minority class, rather than simply replicating existing instances or discarding majority class instances. Later in the second phase, the balanced dataset is divided into training and testing data and then various subsets of training data are used to train the model. Last, the model is tested by using the testing data. The ultimate stage incorporates the weighted fusion approach, which can overcome the limitations of majority voting. According to our research, there has been no research conducted on the diagnosis of clinical diseases using a KNN ensemble classifier.

This highlights a significant gap in the existing literature, as imbalanced datasets are a common challenge in medical diagnosis and can lead to biased or inaccurate predictions. By utilizing multiple diversity structures in a KNN ensemble classifier, there can be a possibility to increase the reliability and accuracy of clinical disease diagnosis, even in the presence of imbalanced data. Our proposition entails a data preprocessing approach that integrates the SMOTE method for data resampling. This technique overcomes the shortcomings of SMOTE by effectively eliminating the noise examples.

2. RELATED WORK:

N. Japkowicz, et al [9] identify that the differences in earlier class probabilities or class imbalances have been reported to hide the performance of standard classifiers. The primary goal in writing this paper is to answer three different questions. Firstly, to understand the nature of the class imbalance problem it is required to understand the relationship between concept complexity, size of the training set and class imbalance level. Secondly, the discussion of basic re-sampling or cost-modifying methods helps us to deal with the class imbalance problem and compare their effectiveness. The results in real-world domains are connected to the results in artificial domains. Lastly, we investigate the assumption that the class imbalance problem does not only affect decision tree systems but also affects other classification systems such as Neural Networks.

Jiang, et al [4] wrote a paper that mainly focuses to forecast the level of risk in a disease. Our platform allows us to combine information from past cases stored in our database and insights from various experts to create a comprehensive medical knowledge base. Using this knowledge base, we can offer relevant recommendations by analyzing and drawing inferences from the information available. The Multi-granularity Linguistic Term Sets (MLTS) framework is employed to tackle the vagueness along with

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

the abstract nature of knowledge by presenting diverse perspectives from experts opinions.

Krawczyk, et al [10] explained that in the literature there are numerous ways to deal with imbalanced datasets. These are generally based on two types of sampling classification. In this paper, for imbalanced classification, an effective ensemble of cost-sensitive decision trees is introduced. To achieve both the selection of classifiers and the assignment of weights for the committee members in the fusion process, we utilized an evolutionary algorithm.

3. METHODOLOGY:

The major health concern is Diabetes, which has affected many people throughout the world. The suggested methodology seeks to create a smart agent that can foretell the presence of any diabetes illness well before any untoward incident occurs. The flow diagram of our proposed system is shown in figure 1.

3.1 Dataset

A Kaggle dataset consisting of around 769 entries. The dataset consists of different features like Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. These all features predict whether a particular person has any Diabetes or not. The target class of these datasets is binary. It depicts whether Diabetes disease is present or not, i.e. 0 or 1.

3.2 Data Preprocessing

Data processing is the process of converting the raw data into a feasible way for implementing a machine learning model on it. This is the first stage to develop a machine-learning model. Its initial raw data may have different odds. So it has to be ensured that the data that is fed to the machine learning model is clean enough. Data preprocessing is done to clean the raw data to acquire desirable results. Data preprocessing includes the following steps: Importing the libraries Bringing the dataset in, Finding the missing information, categorical data encoded, Feature scaling and dividing the dataset into training set and testing set.

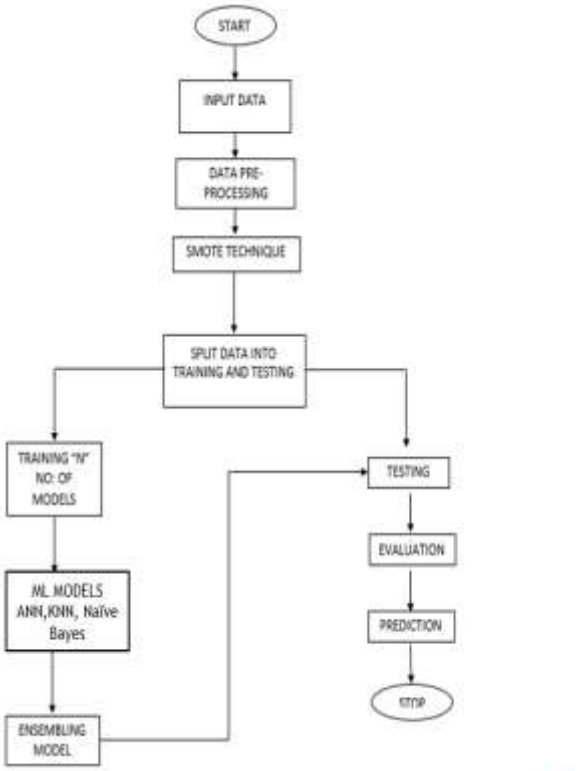


FIGURE 1. Flow Diagram Of Our Proposed System

To process the data, there are many libraries required such as numpy, pandas, random etc. Generally, the datasets are in the form of .csv files. CSV stands for Comma Separated Values. There can be a chance that the data may have missing values. These missing

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

values can lead to mispredictions also. So it is very important to remove the missing data. Missing data can be dealt with through two methods. By removing the specific row, and by figuring out the mean and filling in the missing value.

3.3 Feature extraction

It is today changing into quite common to be operating with datasets of many options. Feature Extraction aims to scale back the number of options in a very dataset by making new options from the present ones. Feature selection aims instead to rank the importance of the present options within the dataset and discard shorter ones. Reducing the number of options to use throughout an applied math analysis will presumably result in many advantages like Accuracy enhancements, Overfitting risk reduction etc. If a lot of options are superimposed than those that are strictly necessary, then our model performance can simply decrease. An optimum number of options must be used.

3.4 Algorithms used

3.4.1 Ensemble Approach

A machine learning method called ensemble learning combines the predictions from various models to produce prognosticative performance.

Ensemble learning consists of three main categories:
• Bagging
• Stacking
• Boosting

Bagging is the process of averaging the results of several decision trees that have been fitted to various samples obtained from the same dataset. Stacking is the process of fitting many models to the same data and using a different model to figure out how to best combine the results. Boosting includes successively adding ensemble members that update predictions made by earlier models, producing a weighted average of the forecasts. In the projected model, the stacking model has been used. The algorithms used are KNN, Naïve Bayes, and ANN.

A Stacked ensemble technique has been used. Depending on the Bayes theorem, the naive Bayes algorithm is one of the supervised learning approaches for handling classification issues. Being a probabilistic classifier, it relies on its judgments of the probability of an item. The main idea behind this algorithm is to make models consecutively and cut back the errors from the previous model. The framework of the stack ensemble learning technique is shown in figure 2.

3.4.2 K-Nearest Neighbors

This algorithm is a straightforward ML technique for both Classification as well as Regression applications. It backed up the idea that the observations in a data collection that are most "similar" to a given piece of information are the observations, and that we should categorize unanticipated points based on the values of the nearby existing points. The range of closest neighbours to utilize is K.

Distance between two points is calculated using Euclidean distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

3.4.3 Artificial Neural Network

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

The term "artificial neural network" originates from the networks of biological neurons that define the architecture of the physical brain of human beings. Dendrites from biological brain networks are used as inputs, organelles as nodes, clumps as weights, and nerve fibres as outputs in artificial neural networks. A computer's ability to see the world and make decisions in a shockingly human manner is enabled by an artificial neural network that attempts to mimic the neural network seen in the human brain.

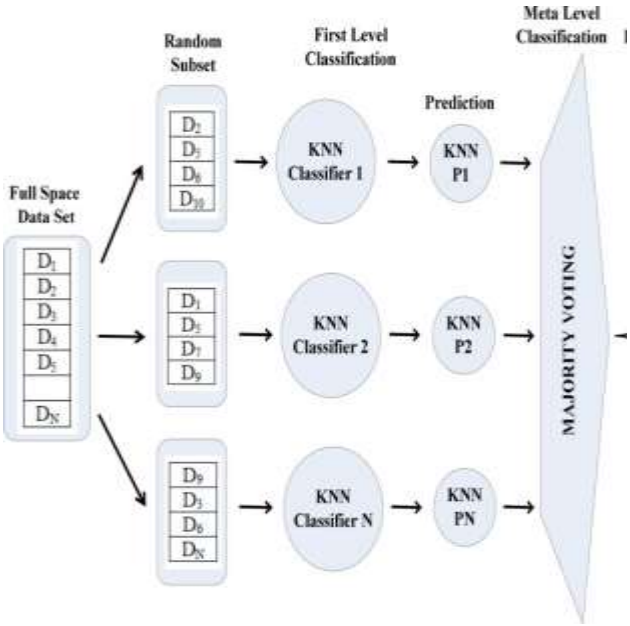


FIGURE 2.The framework of stack Ensemble Learning using different KNN classifiers.

3.4.4 Model Design

The architecture of our proposed system is shown in figure 3.
It involves all the steps from the Registration of the user to getting the results from prediction.

4. RESULTS AND ANALYSIS:

The proposed ensemble model secures higher accuracy of 0.77 when compared to existing models. The dataset consists of 769 entries. Upon the large dataset, the model shows higher accuracy.

MODEL	ACCURACY
Naïve Bayes	0.72
K-Nearest Neighbor(KNN)	0.77
Artificial Neural Network (ANN)	0.50

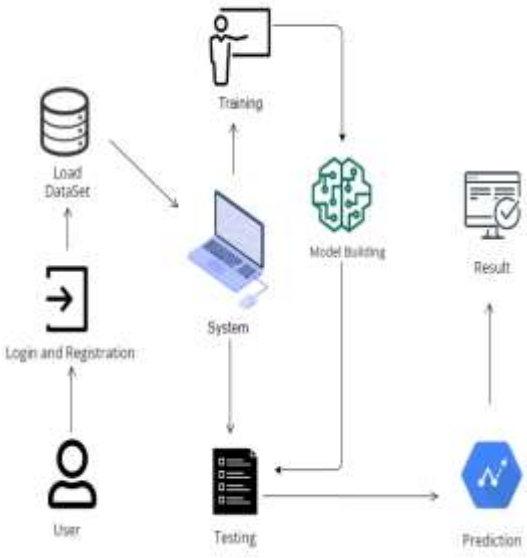
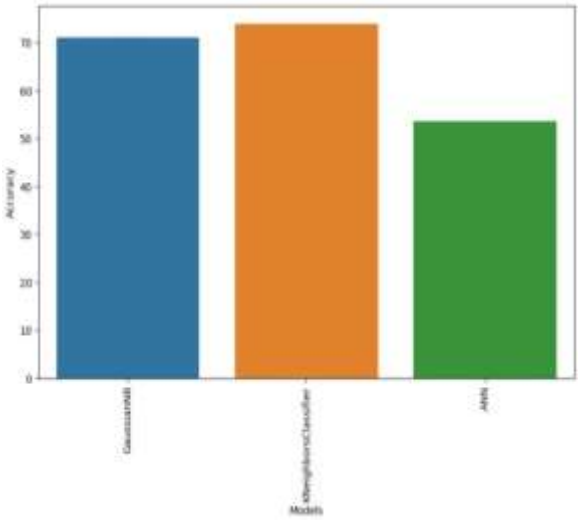


FIGURE 3. Architecture of our proposed system.

The accuracy of the algorithms is shown in the below graph.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023



5. CONCLUSION:

The early identification of Diabetes is particularly helpful in reducing the chances of getting more health issues. When compared to earlier works, the ensemble learning model's accuracy has shown a remarkable improvement. The accuracy shown by the Ensemble approach is 0.77 and Artificial Neural Networks is 0.50. It is ensured that based on the inputs given, the models can predict the presence of Diabetes. The algorithm for artificial neural networks has been used. When compared to these models, earlier research has demonstrated less accuracy.

6. FUTURE SCOPE:

In future, the models' running times and accuracy to improve. To get better results, different hybrid classifiers might be utilized. Without the need for any special equipment, this model can be used when there is a suspicion of any diabetic condition. There can be different novel algorithms be used to obtain desirable results.

REFERENCES:

[1]World Health Organization (WHO), "Cancer".[Online]Available:http://www.who.int/cancer/en/.

[2] Fan, L., Strasserweippl, K., & Li J J et al… Breast cancer in China. The Lancet Oncology, 15(7), e279-e289, 2020.

[3] Wang, K., Makond, B., Chen, K., & Wang, K. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients, Applied Soft Computing, 20 (07), 15-24, 2014.

[4] Jiang, J., Li, X., Zhao, C., Guan, Y., & Yu, Q... Learning and inference in a knowledge-based probabilistic model for medical diagnosis, Knowledge-Based Systems, 139, 58-68, 2017.

[5] Kovalchuk, S V., Krotov, E., Smirnov, P., Nasonov. D A., & Yakovlev, A N. Distributed data-driven platform for urgent decision making in cardiological ambulance control, Future Generation Computer Systems, 79, 144-154, 2018.

[6] Piri, S., Delen, D., & Liu, T. A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets, Decision Support Systems, 106, 15-29, 2018.

[7] Eshtay, M., Hm Faris., & N, Obeid. Improving Extreme Learning Machine by Competitive Swarm Optimization and its application for medical diagnosis problems, Expert Systems with Applications, 104, 134-152, 2018.

[8] Nagarajan, R., & M, Upreti.(2017). An ensemble predictive modelling framework for breast cancer classification, Methods, 131, 128-134

[9]N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," Intelligent Data Analysis, vol. 6, no. 5, pp. 429–450, 2019

[10] Krawczyk., B., M, Woźniak., & G, Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification, Applied Soft Computing, 14, 554-562, 2021.