International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Diabetic Prediction Using Hybrid Smote-Tree Big Data Classification with Artificial Neural Network

[1]Praveenkumar K S, [2]Dr. R Gunasundari
[1]PhD Research Scholar, Dept. of CS, CA & IT, KAHE, Coimbatore
Email: praveen7387@gmail.com
[2]Professor & Head , Dept. of Computer Applications, KAHE, Coimbatore
Email: gunasoundar04@gmail.com

Abstract

Diabetes is one of the worst illnesses now plaguing humanity. The condition is caused by the body's abnormal reaction to insulin, a vital hormone that transforms sugar into the energy required for the normal functioning of daily living. In addition to increasing the chance of developing kidney disease, heart disease, and retinal eye disease, nerve damage, and blood vessel damage, diabetes causes serious consequences in the body. This research provides diabetes prognosis based on hybrid SMOTE-TREE large data categorization utilizing Artificial Neural networks (ANN). Artificial Neural Networks deliver promising results for nonlinear data. Hence ANN is picked for creating the model to predict diabetes among numerous ML (machine learning) techniques. The goal is to develop a decision support system to predict and diagnose diabetes with maximum accuracy, given the parameters. The parameters are set such that the best accuracy is obtained.

Keywords: ANN, Classification, Diabetic Prediction, H-SMOTE

I INTRODUCTION

Diabetes, one of the metabolic illnesses, is the most dangerous disease, posing a significant threat to industrialized and developing countries. High blood glucose content characterizes the condition. The condition is caused by incorrect functioning of the hormone insulin, which is necessary for glucose to enter cells and provide energy to the body [1]. Diabetes is one of the leading worldwide causes of death, especially in developing nations. According to WHO figures, Ebola has wreaked havoc; at least 80 percent or more fatalities occur in low- and middle-income nations, which lack basic and advanced healthcare facilities. India belongs to the group of developing nations, has a large diabetic patient population, and is known as the "diabetes capital of the world" [2-6].

Research conducted by the International Diabetes Federation found that diabetes was responsible for between 1.5 and 5.0 million deaths yearly between 2012 and 2015. The Pima Indian diabetes dataset (PIMA) [7] is used to diagnose patients, and it contains characteristics (symptoms) that help classify them into two categories. Since ANNs [8] are effective at solving a wide variety of problems, including function approximation (for mathematical functions or formulas), prediction (machine learning), curve-fitting, classification (machine learning), and clustering (unsupervised learning), the term "universal approximator" has been coined to describe them. Using ANN, we classify diabetes cases from the PIMA dataset.

ANN may be used to create and apply these models, which have shown to be more useful, efficient, and successful in various medical domains, including analysis, diagnosis, and prediction, and which aid both experts and the general public [9]. ANN's mathematical description of the human brain system indicates that training and generalization are successful. Nonlinear functions provide the basis of most ANN approaches, in which the relationship or correlation between input data is ambiguous or difficult. An ANN stacks numerous layers of nodes also called neurons. Each neuron in a traditional statistical model of an ANN is directly connected to the neurons in the upper layers through a weighted value representing the connection's strength or power [10-14].

The input of each neuron is influenced by the weighted permutation of many input signals, which may include separate calculations, followed by the neuron's output. By applying the transfer function to the weighted inputs of neurons, individual threshold values may be obtained. The activation function transfers the information to the next cell if the threshold value is exceeded. It is crucial to comprehend when ANNs give a prediction,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

perceptual categorization, pattern recognition, and training based on their functions [15]. Artificial intelligence, robotics, image processing, and other cutting-edge technologies rely heavily on a wide variety of contemporary ANN models, such as deep learning models, recurrent neural networks, and genetic algorithms, with special emphasis on their use in the healthcare sector. Stroke, bone densitometry, hepatitis B, and breast cancer are among conditions that ANN excels in identifying, evaluating, and predicting. [16]. Using a dataset of female patients with similar features, this study provides a prediction framework for an accurate diabetes diagnosis. Importantly, the problem was resolved using a typical regression model. The main contributions of this study are diabetic prediction using hybrid SMOTE-TREE with ANN algorithm.

The subsequent sections are organized as follows. In the second part, previous research on diabetes prediction is reviewed. Section 3 of the proposed procedure details the H-SMOTE tree-based architecture for diabetes prediction in its entirety. Section 4 discusses the experiments and results, while Section 5 presents the conclusion and suggestions for further study.

II BACKGROUND STUDY

Chatragadda, B. et al. [3] this research aims to evaluate diabetes therapy in the health care business using massive data analysis. The description of diabetes treatment's future research plan may provide cutting-edge data and analysis that provide the most effective medical results. Using Spark, the author predicted which sexual orientation and race had a larger likelihood of being afflicted with diabetes. Apache Spark may also minimize the processing time necessary for genomic sequencing.

Ding, S. et al. [4] Because diabetes and its complications may cause such significant damage to human health, accurate forecasts of diabetic outcomes are necessary for avoiding and treating diabetic issues and increasing the survival rate of diabetic patients. This work introduces a unique approach to predicting diabetes complications based on a similarity-enhanced latent Dirichlet allocation (seLDA) model. Author performs data preprocessing, then uses similarity estimates between pairs of medical records to guide the topic mining process using seLDA for diabetes complications. The author uses the latent topics of the progress notes to build the vector space for the progress notes. To improve the prediction model for the multi-label classification issue, support vector machines were used.

Fiarni, C et al. [6] Using clustering and classification data mining methods and their corresponding algorithms, a prediction model for complications of diabetes was developed. The approach classifies diabetes-related medical information into four categories: Nephropathy, Retinopathy, neuropathy, and mixed consequences (other). The author evaluates the effectiveness of clustering and classifying strategies to develop the most effective rule-based model for prediction. Classification gives improved information, performance, and the capacity to group characteristics and sub-features into three major microvascular diabetic problems compared to clustering. The author may uncover the most important risk factor for each diabetic complication condition through data mining. Nephropathy was more sensitive to this illness, even though the blood glucose level and duration of diabetes were troublesome. The study indicates that glucose levels and genes (diabetes in the family) do not affect the development of some diabetic problems. Also, the author discovered that blood pressure in the hypertension crisis range was the most frequent risk factor for Retinopathy.

Islam, M. M et al. [9] these authors' research demonstrates that deep learning (DL) may play a significant role in diagnosing referable Diabetic Retinopathy (DR) with excellent sensitivity, specificity, and reproducibility. Future deployment of an automated system based on DL might potentially modify the diagnosis procedure for DR. Automated methods may improve DR screening quality, broaden patient access to treatment, and reduce the financial burden of the process. If caught and treated early enough, this problem may not ever start.

Prasad, S et al. [12] big data analytics in Hadoop implementation provide a systematic approach to achieve health results equal to the accessibility and affordability of healthcare for all citizens. This project is designed mainly for use by rural and urban communities.

Sujatha, V. et al. [15] Given that diabetes has risen to prominence as a major health problem, addressing it has become a top priority for scientists and healthcare administrators. Prediction models that include patient treatment data and prognostic outcome findings may be developed using the data sets obtained by statistical or advanced pattern recognition algorithms. It's possible to utilize this data for diabetes monitoring and clinical decision support to better care for patients. Given the seriousness of the ailment and the urgency with which it must be treated, the author places more emphasis on actual results than on predicting models. Over the last decade, researchers have created a variety of different prediction models for dealing with diabetes and its complications. Prediction models were created using either multiple logistic regressions or a linear regression with the same benefits.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

III MATERIALS AND METHODS

Diabetes may affect individuals of all ages. These characteristics may change according to age, gender, lifestyle, glucose and insulin levels, blood pressure, and other variables. Artificial neural networks (ANNs), support vector machines (SVMs) combined with fuzzy logic (FL) is only some of the methodologies used to make diabetes diagnostic predictions today. These techniques involve processing time, accuracy tradeoffs, and information extraction from concealed data. Finding the ideal answer to complex problems may not need a straightforward technique. The following graphic depicts how ANN algorithms predict, verify, and routinely test the network to improve self-reliance and substantial assurance. ANN training generates operations and parameters, which are then compared with predicted and realized ANN and H-SMOTE Tree values.

a) Dataset:-

Dataset Taken from: https://www.kaggle.com/uciml/pima-indians-diabetes-database

Parameters

1. Number of times pregnant,
2. The concentration of plasma glucose rate,
3. Blood pressure(mm Hg),
4. Triceps skinfold thickness(mm),
5. Serum insulin amount(mu U/ml),
6. Body mass index,
7. Diabetes pedigree,
8. Age in years,
9. Nature of Exercises- MET value

Pregnanci	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1

Figure 1: Dataset Structure

Figure 1 denotes the dataset structure with attribute values of diabetic type ii.

3.1 AMALGAM MULTIVARIATE STATISTICAL MODELING ALGORITHM [AMSM]

1: A sliding state with the size of s_1, s_2 moves data by data through the entire dataset and computes the manner value in every underlying with PCA.

2: p1 and p2 are the standard deviations for intensities that are better and worse than the manner worth in each underlying condition. These two values should be assigned to the corresponding state's center information.

3: Multivariate the data by separating each data intensity by its matching std.

4: Apply the AMSM algorithm in ANN.

Main loop: for j = 1; …. J

i) Clustering step:

a) For each s1, s2 sliding reference patch, identify KNN-like patches with a state size of R, R around the reference patch. The reference patch and its contiguous neighbors form one cluster.

b) Approximation of Gaussian parameters of every cluster.

ii) Denoising step: Denoise every cluster's patches by the active filter.

iii) Acquire reconstructed diabetic data by a weighted average of de-noised patches.

5: Reconstruct the concluding diabetic data by applying the inverse of Multivariate in step 3. It is proficient in multiplying the matching std of every patch into its middle data strength.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

3.2 Support Vector Machine

Support Vector Machine (SVM) offers cutting-edge performance in real-world applications and is effective with unobserved data. In contrast to neural networks, SVM yields repeatable results. The calculation of error boundaries facilitates model generalization. SVM operates by transforming the input vector into high-dimensional feature space and maximizing the margin to find the optimal hyperplane that divides classes. Figure 2 illustrates the SVM algorithm. Support vectors are the chosen subset of training data points. A kernel approach that provides a mapping to a high-dimensional space is utilized to map the input space to the feature space. Radial basis and linear kernels are the two most often used kernels in support vector machines (SVM). SVMs were initially envisioned as binary classifiers; hence, various strategies are used to extend SVM to multiclass settings. The expression represents the SVM formula for estimating the decision function from the training dataset.

$f(x) = sign(\sum_{n=1}^{l} y_n \propto_n . k(x, x_n + b))$ -------(1)

Here, 7 represents the total number of support vectors, b represents the bias factor, and - 1,+1 represents the class sign of the support vector to which the test sample belongs. Solving the following quadratic optimization problem yields the answer:

$min \frac{1}{2} w^T w + C \sum_{i=1}^{p} \varepsilon_i$-----(2)

subject to

$y_i(w^T \emptyset(x_i) + b) \geq 1 - \varepsilon_i$ ------(3)

$\epsilon_i \geq 0, i = 1, \ldots, p$-------(4)

There must be more support vectors than data points in the dataset. It is difficult for SVM to handle dynamic systems. This may be remedied by transforming the time series to fixed-length vectors before SVM analysis

.

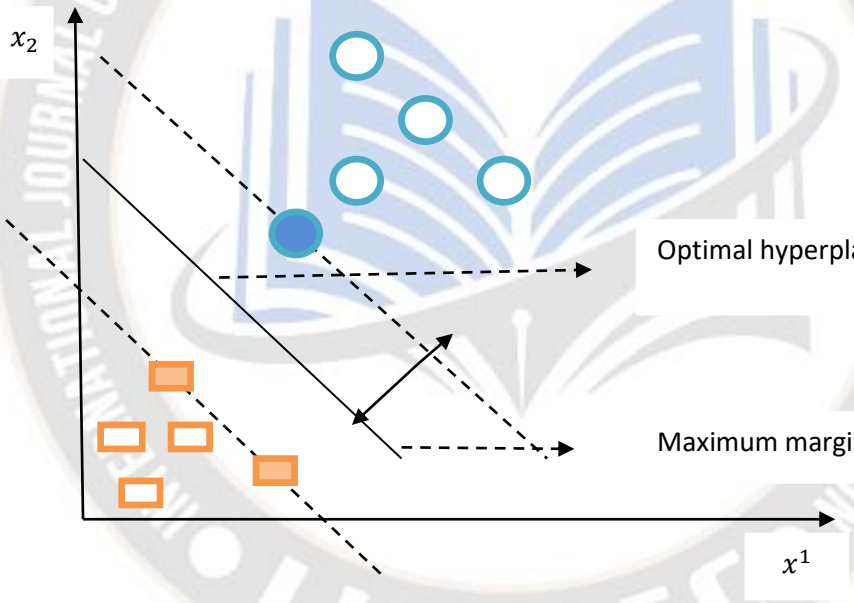

Figure 2: Support vector machine

3.3 ANN

Neural Networks (NNs) are a kind of computer model that are meant to mimic the functioning of neurons in the human brain. When ANNs include hidden layers, the calculation is parallelized, resulting to a significant increase in processing speed. These models may be used to simulate real-world applications like image recognition, voice recognition, and text analysis. For ANNs, the activation function stands in for a neuron's action potential and causes a node to fire if the input signal's strength rises beyond a certain threshold. It is possible that this activation is nonlinear, providing ANNs the ability to effect nonlinear changes. With the right architecture and activation function, ANNs may represent high-dimensional transformations from the input space to the output space. The architecture of ANN is shown in figure 3.

The general architecture of ANN is as follows: -

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

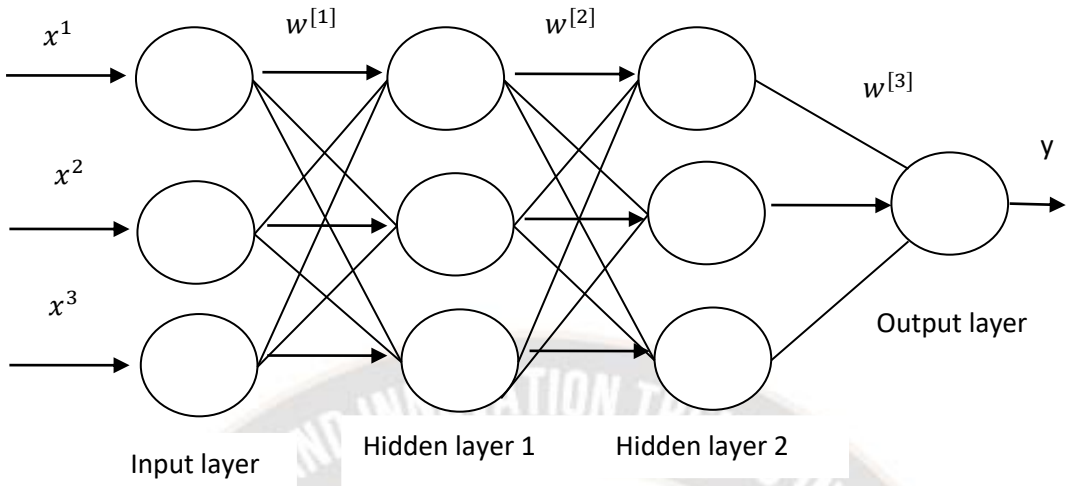

Figure 3: Artificial neural networks

The first layer is known as the input layer, the second as the hidden layer, and the third as the output layer in this context. Using the equation, each node collects inputs (1),

$f(x) = w_0 + \sum_{i=1}^{n} w_i x_i$------(5)

For input, each node employs a sigmoid activation function,

$g(x) = \frac{1}{1+e^{-x}}$---------(6)

A neuron's structure comprises a summation junction and an activation function. Figure 2 displays a neuron that receives N inputs, x1, x2,...xM. Each connection between these inputs and the neuron is represented by a weight, denoted by w1, w2,..., we. At the summation junction, their inputs are multiplied by their respective weights and added. Net output is the output of the summing junction. Consequently, this component performs multiplication and addition operations (Multiply-Accumulate, MAC). This is one of the essential calculations performed by hardware implementation of artificial neural networks.

$a = \sum_{i=1}^{M} x_i w_i$-----(7)

The activation function 'a is then implemented. Using a nonlinear activation function [4, the net value is translated to output. The function may be linear, exponential, or nonlinear. The threshold function is what decides whether or not a neuron will fire. The most popular activation function is the sigmoid function. Therefore, neuron output may be described as follows:

$y = f(a)$-------(8)

The sigmoid activation is as follows.

$f(a) = \frac{1}{1+e^{-a}}$---------(9)

Computing activation functions are another key aspect of the hardware implementation of artificial neural networks [13]. It determines whether a neuron is active or not. Nonlinear functions, such as the sigmoid function, are often utilized because they may approximate any function. Owing to the impossibility of implementing the sigmoid function directly due to its infinite exponential series, approximate methods are used instead. The sigmoid function approximation technique influences the precision of the activation functions and, by extension, the training of the whole ANN. The most common technique for approximation is the Taylor series approach's piece-wise linear approximation. The sigmoid function was approximated using linear curves of the form y = axe + b. This approximation permits the expression of the sigmoid function as follows: [12].

$$f(v) = \begin{cases} 0 & ,for\ v > 4 \\ 0.0625x + 0.75, for 0 < v < 4 \\ 0.5 & ,for\ v = 0 \\ 0.0625x + 0.25, for -4 < v < 0 \\ 1 & ,for\ v < -4 \end{cases}$$ ----------(10)

3.4 HYBRID CLASSIFICATION APPROACH OF SMOTE, RESAMPLE AND ATTRIBUTE SELECTION TECHNIQUES

A hybrid classification strategy is developed in this work to overcome the limitations that a single learning model or statistical technique cannot solve. No classification model can handle all forms of data properly. As a result, the goal is to integrate many learning models or statistical models to improve performance by decreasing their restrictions and enhancing their different techniques. A hybrid intelligent system has numerous levels, and each

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

level provides new information to the next. As a result, the precise functionality at all levels adds to the total functioning of the model. This hybrid classification approach is inspired by two sampling strategies to balance the dataset, SMOTE and resample. The attribute selection approach is used for this well-balanced dataset. Both sampling strategies are efficient and work in different ways. When constructing synthetic instances, SMOTE does not consider neighboring instances from different classes, resulting in overlapping and noise. If undersampling is considered in resampling, it discards potentially beneficial information that might be critical, and it raises the risk of over-fitting since it repeats the uncommon class occurrences. As a result, combining these two sampling procedures maximizes their advantages while minimizing their negatives. The hybrid sampling technique functions at the instance level of balanced data, while the attribute selection method operates at the feature level. It automatically discards less valuable attributes, decreasing the dimension for better categorization predictions. This well-balanced and dimension-reduced dataset is now being classified. The proposed technique is validated in this research using a k NN classification model. As a result, the dataset altered by the proposed technique is now classified using k NN. Nonetheless, this strategy works well with other classification models as well. The proposed architecture is shown in figure 4.

The Basic Steps of the Proposed H-SMOTE Classification Method

Preprocess the dataset by loading it (Bi).

2. Create a hybrid classification technique using SMOTE, Resample, and Attribute Selection (H-SMOTE)

Obtain a converted dataset (BI) with the desired properties.

4. Make use of the (BI) k NN classifier.

5. Output: a balanced dataset with selected attributes and k NN classification results.

Figure 4.2 shows how an imbalanced dataset is preprocessed using the hybrid classification technique H-SMOTE and transformed into a balanced dataset with defined features before being supplied to the classifier for classification.

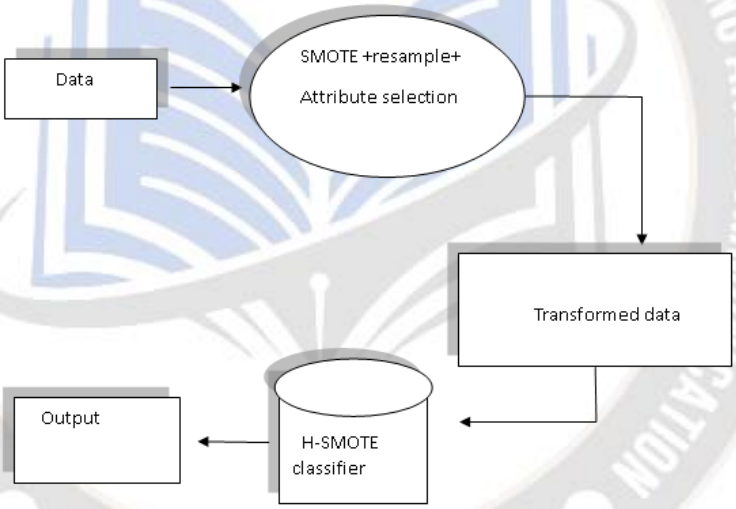

Figure 4: The Proposed Hybrid Classification Technique (H-SMOTE)

The suggested classification strategy is detailed in the following algorithm H-SMOTE.

The H-SMOTE Classification Method, a Proposed Algorithm

1. T: Training Set, M: SMOTE d sample, N: SMOTE, and Resample percentage are all inputs.
2. Output: R set of selected samples
3. Begin
4. M - α;
5. s_1 ←MinClassInstances (T); # minority class instances
6. s_2 ←MajClassInstances (T); # minority class instances
7. UndSampleMaj← false;
8. M ←SMOTE (s_1 , N, UndSampleMaj) # call SMOTE method
9. M← $M \cup s_2$;
10. With Replacement ←True;
11. Resample (M, N, with Replacement) # call resample method
12.R_1 ← Resampled MinInstance;
13. R_2 ← Resampled MajInstances;
14. $R \leftarrow R_1 \cup R_2$;
15. $R \leftarrow R \cup M$;

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

16. R ←AttributeSelection(R) # call attribute selection method
17. end
18. return R

the transformed dataset R is classified using the k NN algorithm.

IV RESULTS AND DISCUSSION

4.1 Performance metrics

There are four possible outcomes for a single prediction: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (FN). True Positive and True Negative are both acceptable categories. The categories are shown using a Confusion Matrix. True Positive describes a sample that is both positive and projected to be positive. False Positive occurs when a sample believed to be positive is negative. True Negative occurs when both the sample and the projection are negative. False Negative refers to a sample that should be negative but is positive.

For this study, we used: 1. Accuracy. 2. Precision. 3. Recall. 4. F-measure; equations for evaluation and analysis

$$Accuracy = \frac{TP+}{TP+TN+FP+FN} \text{ ---------- (11)}$$

$$Precision = \frac{TP}{TP+FP} \text{ ---------- (12)}$$

$$Recall = \frac{TP}{TP+FN} \text{ ---------- (13)}$$

$$Fmeasure = \frac{2.Precision\ x\ recall}{Precision+Recall} \text{ ---------- (14)}$$

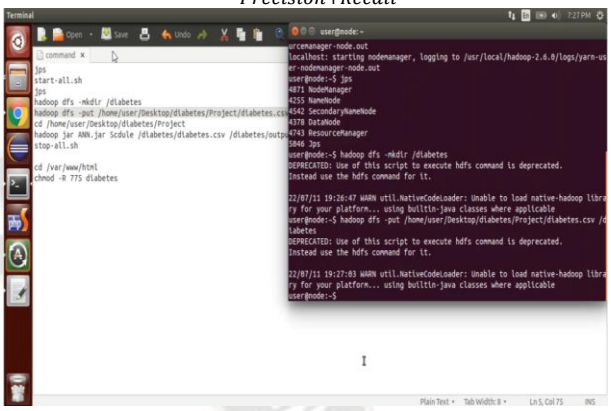

Figure 5: Implementing Diabetic Prediction in Bigdata window

By implementing java with hadoop the implementation screen is shown in figure 5.

Table 1: ANN Training and testing loss and accuracy comparison table

Sno	Training Accuracy	Training loss	Testing accuracy	Testing loss
1	0.4	1.75	0.3	2
2	0.5	1.66	0.4	1.66
3	0.7	1.45	0.5	1.45
4	0.75	1.32	0.65	1.32
5	0.8	0.78	0.7	0.78
6	0.85	0.52	0.75	0.52
7	0.95	0.21	0.9	0.21

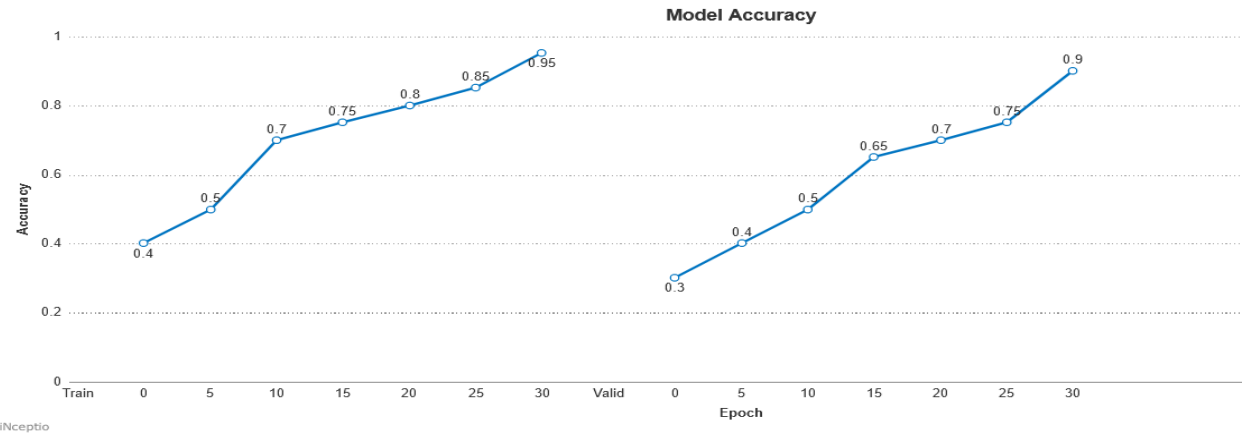

Figure 6: ANN Training and validation accuracy

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

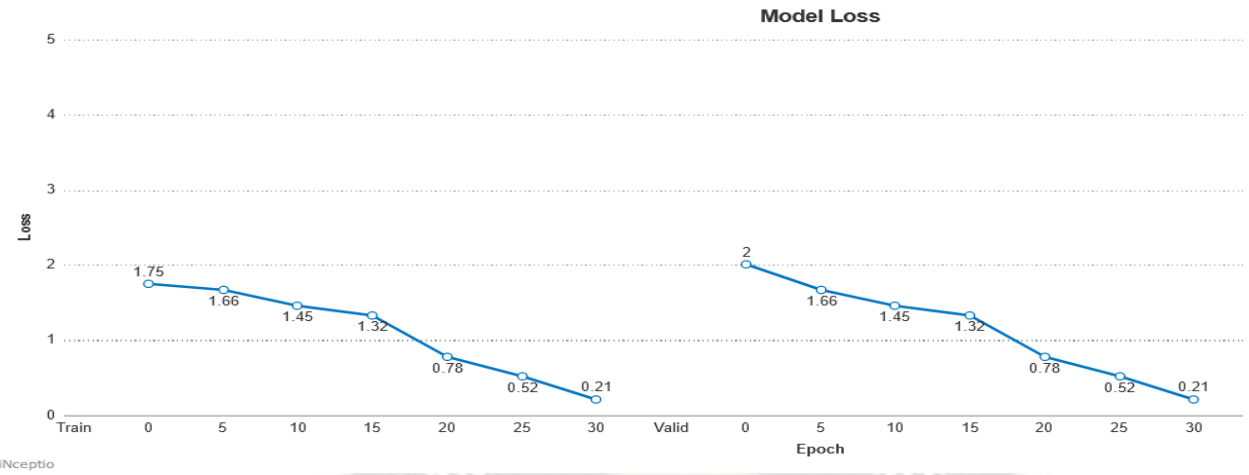

Figure 7: Training and testing loss values

Training and testing values of ANN is shown in table 1. The training and testing accuracy comparison chart is shown in figure 6. X-axis denotes the training and validation epoch number. Y-axis denotes the accuracy. The loss values of training and validation is shown in figure 7.

Table 2: Classification of performance metrics

Sno	Algorithm	Accuracy	Precision	Re-call	F-measure
1	SVM	0.93	0.8	0.8	0.8
2	ANN	0.8	0.8	0.7	0.7
3	H-SMOTE	0.99	0.8	0.9	0.9

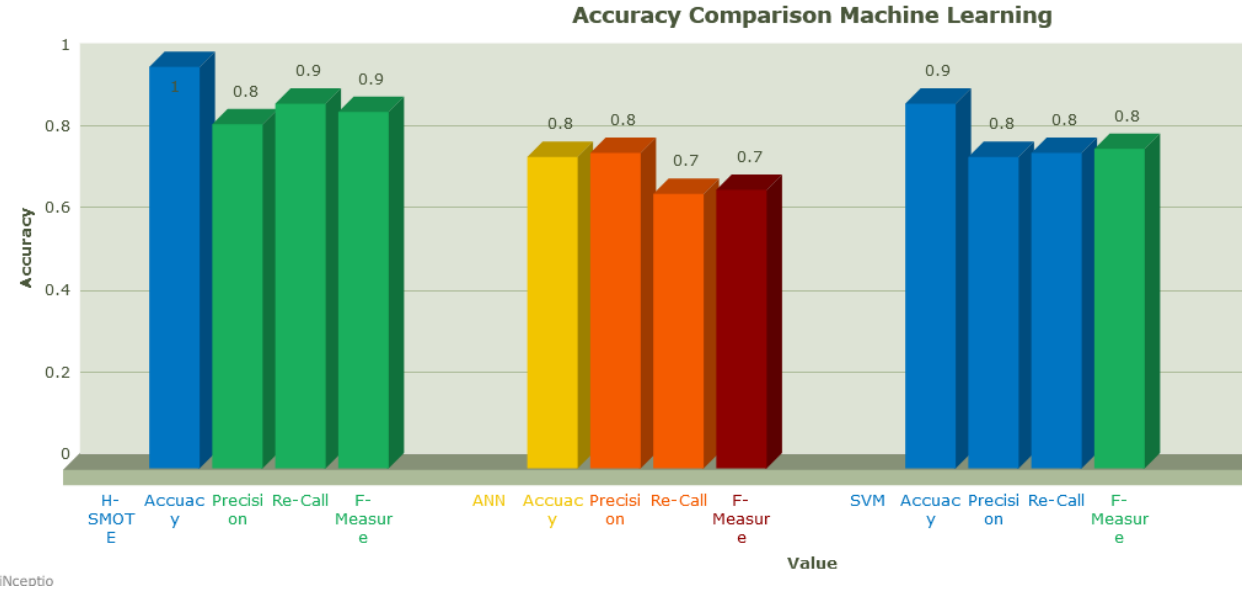

Figure 8: Classification metrics comparison chart

The classification performance metrics like accuracy, precision, recall and f1measure. The algorithm like ANN, SVM and H-SMOTE performance metrics is shown in table 2. Figure 8 represents classification comparison chart. X-axis denotes classification metrics and Y-axis denotes accuracy in percentage.

V. CONCLUSION

As a result of the significant improvement in expert systems and machine learning techniques, their influence is permeating more and more application fields every day. The medical industry is not an exception. Sometimes, making decisions in the medical industry may be quite difficult. Classification systems used to make medical decisions are given with medical data, which they analyze more thoroughly in less time. In this research investigation, an H-SMOTE tree with ANN Classification was proposed. We used this approach to diagnose diabetes and assessed the accurate learning methods. For further we suggest to develop and monitoring diabetes using android application.

REFERENCE

[1] Ameena, R. R., & Ashadevi, B. (2020). Predictive analysis of diabetic women patients using R. Systems Simulation and Modeling for Cloud Computing and Big Data Applications, 99–113. doi:10.1016/b978-0-12-819779-0.00006-x

[2] Cao, P., Ren, F., Wan, C., Yang, J., & Zaiane, O. (2018). Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis. Computerized Medical Imaging and Graphics. doi:10.1016/j.compmedimag.2018.08.008

[3] Chatragadda, B., Kattula, S., & Guthikonda, G. (2018). Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data. 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). https://doi.org/10.1109/rteict42901.2018.9012339

[4] Ding, S., Li, Z., Liu, X., Huang, H., & Yang, S. (2019). Diabetic Complication Prediction Using a Similarity-Enhanced Latent Dirichlet Allocation Model. Information Sciences. doi:10.1016/j.ins.2019.05.037

[5] Esteban, S., Rodríguez Tablado, M., Peper, F. E., Mahumud, Y. S., Ricci, R. I., Kopitowski, K. S., & Terrasa, S. A. (2017). Development and validation of various phenotyping algorithms for Diabetes Mellitus using data from electronic health records. Computer Methods and Programs in Biomedicine, 152, 53–70. doi:10.1016/j.cmpb.2017.09.009

[6] Fiarni, C., Sipayung, E. M., & Maemunah, S. (2019). Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm. Procedia Computer Science, 161, 449–457. doi:10.1016/j.procs.2019.11.144

[7] Hassan, S., Dhali, M., Zaman, F., & Tanveer, M. (2021). Big data and predictive analytics in healthcare in Bangladesh: regulatory challenges. Heliyon, 7(6), e07179. doi:10.1016/j.heliyon.2021.e07179

[8] Huang, F., Abbasi-Sureshjani, S., Zhang, J., Bekkers, E. J., Dashtbozorg, B., & ter Haar Romeny, B. M. (2019). Vascular biomarkers for diabetes and diabetic retinopathy screening. Computational Retinal Image Analysis, 319–352. doi:10.1016/b978-0-08-102816-2.00017-4

[9] Islam, M. M., Yang, H.-C., Poly, T. N., Jian, W.-S., & Li, Y.-C. (Jack). (2020). Deep Learning Algorithms for Detection of Diabetic Retinopathy in Retinal Fundus Photographs: A Systematic Review and Meta-Analysis. Computer Methods and Programs in Biomedicine, 105320. doi:10.1016/j.cmpb.2020.105320

[10] Khanna, N. N., Jamthikar, A. D., Gupta, D., Nicolaides, A., Araki, T., Saba, L., … Suri, J. S. (2019). Performance evaluation of 10-year ultrasound image-based stroke/cardiovascular (CV) risk calculator by comparing against ten conventional CV risk calculators: A diabetic study. Computers in Biology and Medicine, 105, 125–143. doi:10.1016/j.compbiomed.2019.01.002

[11] Nagarathna, R., Tyagi, R., Battu, P., Singh, A., Anand, A., & Ramarao Nagendra, H. (2020). Assessment of Risk of Diabetes by using Indian Diabetic Risk Score (IDRS) in Indian population. Diabetes Research and Clinical Practice, 108088. doi:10.1016/j.diabres.2020.108088

[12] Prasad, S. T., Sangavi, S., Deepa, A., Sairabanu, F., & Ragasudha, R. (2017). Diabetic data analysis in big data with predictive method. 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET). doi:10.1109/icammaet.2017.8186738

[13] Rastogi, R., Singhal, P., Chaturvedi, D. K., Satya, S., Arora, N., Gupta, M., & Saxena, M. (2021). Study of asian diabetic subjects based on gender, age, and insulin parameters: healthcare application with IoT and Big Data. Healthcare Paradigms in the Internet of Things Ecosystem, 333–362. doi:10.1016/b978-0-12-819664-9.00015-6

[14] Ruan, X., Li, Y., Jin, X., Deng, P., Xu, J., Li, N., … Xu, L. (2021). Health-adjusted life expectancy (HALE) in Chongqing, China, 2017: An artificial intelligence and big data method estimating the burden of disease at city level. The Lancet Regional Health -

Western Pacific, 9, 100110. doi:10.1016/j.lanwpc.2021.100110

[15] Sujatha, V., Prasanna Devi, S., Vinu Kiran, S., & Manivannan, S. (2016). Bigdata Analytics on Diabetic Retinopathy Study (DRS) on Real-time Data Set Identifying Survival Time and Length of Stay. Procedia Computer Science, 87, 227–232. doi:10.1016/j.procs.2016.05.153

[16] Yang, Y., Li, Y., Chen, R., Zheng, J., Cai, Y., & Fortino, G. (2021). Risk Prediction of Renal Failure for Chronic Disease Population Based on Electronic Health Record Big Data. Big Data Research, 25, 100234. doi:10.1016/j.bdr.2021.100234