JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2758 - 2764
https://publishoa.com
ISSN: 1309-3452

Cancer Prediction using Machine Learning: A Survey

Soumya K N[1], Dr.Rajapraveen.K.N[2]

[1] *Research scholar, School of CSE, Jain University, Bengaluru, Karnataka, India*
[2] *Associate professor, School of CSE, Jain University, Bengaluru, Karnataka, India*

Email: [1] *knsoumya@jainuniversity.ac.in,* [2] *p.raja@jainuniversity.ac.in*

ABSTRACT

Carcinoma is a terrible illness that has been rising in morbidity at an alarming rate in recent years all over the world. Computer-assisted cancer prediction has made significant progress thanks to the rapid growth of computer science and machine learning technology. A novel technique that includes many machine learning models and uses deep learning in an ensemble manner. Five distinct categorization models using meaningful gene data are been derived from differential gene expression analyses. The results of the five classifications are then combined using a deep learning algorithm. To choose features in T2DM, Fisher's score, RFE, and a decision tree are been used. The prevalence of diabetes was predicted using random forest, logistic regression, SVM, and MLP. The MIMICIII data collection is being used to construct and train several algorithms aimed at predicting DM patient death by using deep learning model. Breast cancer (BC) is also one of the most common reasons of worry across the world. It was the world's second most often diagnosed cancer and the fifth leading cause of mortality. High precision outcomes are frequently obtained at the expense of sensitivity. Distance-based clustering with Euclidean distance, the k-means method, and discretization are among the machine learning techniques employed.

Keywords— Cancer Prediction, Machine Learning Technology, Breast Cancer, Computer-Assisted Cancer Prediction

I. INTRODUCTION

Carcinoma is a terrible illness that has been rising in morbidity at an alarming rate in recent years all over the world. Computer-assisted cancer prediction has made significant progress thanks to the rapid growth of computer science and machine learning technology [1]. A novel technique that includes many machine learning models and uses deep learning in an ensemble manner [12] Five distinct categorization models using meaningful gene data are been derived from differential gene expression analyses. The results of the five classifications are then combined using a deep learning algorithm [2][9]

Assist is much needed for the medical practitioners in detecting Type 2 diabetes early and correctly diagnosing it.[3][6] For boosting the accuracy of diabetes prediction, researchers used bioinformatics theory and supervised machine learning techniques.

To choose features in T2DM, Fisher's score, RFE, and a decision tree are been used. The prevalence of diabetes was predicted using random forest, logistic regression, SVM, and MLP. In the community, ML models for T2DM prediction perform well. Before they can be deployed at scale, they need to improve their methodology, reporting, and validation.[7][8][10]

The MIMICIII data collection is being used to construct and train several algorithms aimed at predicting DM patient death by using deep learning model.[11]

Breast cancer (BC) is also one of the most common reasons of worry across the world. It was the world's second most often diagnosed cancer and the fifth leading cause of mortality. High precision outcomes are frequently obtained at the expense of sensitivity. Distance-based clustering with Euclidean distance, the k-means method, and discretization are among the machine learning techniques employed.[4]

II. LITERATURE SURVEY

Xia, Chao, et al. "A convolutional neural network based ensemble method for cancer prediction using DNA methylation data." *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*. 2019.

[1] The author introduces a convolutional neural network based multi-model ensemble method for cancer prediction using DNA methylation data. The author has chosen five basic machine learning methods as the first stage classifiers and conduct prediction individually. Then, a convolutional neural network is used to find the high-level features among the classifiers and gives a credible prediction result.

Author has introduced a convolutional neural network-based ensemble method for cancer prediction using DNA methylation data. Author initially performed t-test to choose a set of significantly differential methylation points. Then, the selected feature was feed into Naive Bayesian Classifier, k-Nearest Neighbor, Decision Tree, Random Forest and

Gradient Boosting Decision Tree five basic classifiers for the first stage classification. Machine learning methods are used to establish a classifier for DNA Methylation 450K datasets to distinguish cancer from normal samples
LUAD dataset contains 395995 features where 460 were tumor samples and 32 were normal samples, similarly LIHC Dataset contains 395911 datasets where 379 were tumor samples and 50 were normal samples and KIRC dataset contans395808 features where 320 were tumor sample and 160 were normal samples.
Prediction results of Naive Bayesian of LUDA is 98.98%, LIHC is 98.37%,KIRC of 98.54% , k-nearest neighbor of LUDA is 98.37%, LIHC is 98.37%, KIRC of 98.54% , decision tree of LUDA is 98.917%, LIHC is 96.74%,KIRC of 98.96%, random forest of LUDA is 97.36%, LIHC is 92.31%,KIRC of 95.83% and gradient boosting decision tree of LUDA is 97.915%, LIHC is 97.44%,KIRC of 99.17% methods on three datasets.
Experimental results on three DNA methylation datasets of Lung Adenocarcinoma, Liver Hepatocellular Carcinoma and Kidney Clear Cell Carcinoma show the proposed ensemble method can uncover the intricate relationship among the classifiers automatically and achieve better performances.
The above experimental results indicate that the convolutional neural network based multi-model ensemble method can learn the intricate relationship among the classifiers automatically and achieve better prediction results.
[2] Carcinoma is one of the complicated global health issue with a high fatality rate. As Progress in cancer prediction has been increasingly made based on gene expression, providing insight into effective and accurate treatment decision-making, thanks to the rapid development of increased sequencing technology and the application of various machine learning methods that have developed in recent years. Therefore, creating machine learning approaches that can accurately discriminate malignancy patients from healthy ones is a hot topic right now. Nevertheless, yet no classification system has yet to outperform all others in the field of cancer prediction.An innovative technique is been introduced that includes numerous machine learning models and uses deep learning in an ensemble approach. He has provided five distinct categorization models using meaningful gene data derived from differential gene expression analyses. The outputs of the five classifiers are then combined using a deep learning algorithm.
To choose relevant genes for downstream classifications, the author used the DESeq approach. The DESeq approach is commonly used to determine if an observed change in read count for a given gene is noteworthy, such that, whether it is higher than what might be anticipated just from natural random variation. The substantially differentially expressed genes are screened and chosen in differential expression analysis by setting the BH-adjusted p-value and fold change level criteria.
Raw count data and normalized fragments per kilobase per million (FPKM) data are both used in the technique. The normalized FPKM data were utilized in the classification and ensemble method after the raw count data were used to determine the substantially differentially expressed genes. The data set of LUDA which has around 20532 genes, around 125 tumor samples were identified 36 were normal. Similarly, STAD and BRCA which has 29699 genes, around 238 tumor samples were identified, 33 were normal. The precision, Recall and Accuracy for LUDA is 98.46%,97.37% and 96.80 respectively, STDA had precision of 99.42%, Recall of 97.22% and Accuracy of 96.59% and BRCA has precision of 97.77%, Recall of 97.42% and Accuracy of 95.76%.
Three public RNAseq data sets of three types of malignancies, Lung Adenocarcinoma, Stomach Adenocarcinoma, and Breast Invasive Carcinoma, were used to evaluate the proposed deep learning-based multi-model ensemble technique. When compared to utilizing a single classifier or the majority voting approach, the test results show that it improves cancer prediction accuracy for all of the evaluated RNA-seq data sets.
The proposed deep learning-based multi-model ensemble method for cancer prediction has been found to be reliable and effective by taking full use of several classifiers.
[3] The goal of this study article is to assist medical practitioners in detecting Type 2 diabetes early and correctly diagnosing it. Based on eight clinical parameters included in the widely known PIMA dataset, Author has employed bioinformatics theory and supervised machine learning techniques to improve the accuracy of diabetes prediction.
The PIMA dataset has 768 occurrences and is made up of a diverse group of diabetic and non-diabetic individuals. Patients with diabetes account for 34.9 percent of the total population, whereas non-diabetic patients account for 65.1 percent.
The harmonic mean of accuracy and recall is the F1 score. It informs us how precise (or robust) the classifier is (how many times it properly classifies) (it does not miss a significant number of instances). Because the dataset was uneven, with an unequal proportion of good and bad outcomes, the authors were compelled to utilize it.
The F1 score rises by 0.029 when the areas of BloodPressure and Pregnancies are eliminated from the test and training sets, according to the author (from 0.824 to 0.853). However, in order to avoid overfitting, we didn't remove any columns from our data frame. Another interesting finding was that the gradient boosting technique prioritized the SkinThickness value above the BloodPressure attribute.
The model's efficacy is demonstrated by its F1 score of 0.853 and out-of-sample prediction of 89.94 percent. For the problem of diabetes diagnosis, the performance and assessment of the employed machine learning approaches were carefully explored.
Hyperglycemia must be detected early on in order to be treated effectively.
Author employs artificial intelligence in bioinformatics to uncover information and forecast future events by converting

clinical data into meaningful outcomes.

[4] The study assesses the effectiveness of machine learning approaches for predicting breast cancer recurrence. The datasets may be publicly accessible (e.g., online) or they may be the outcome of a partnership between institutions and research teams that is not open to the public. Feature selection can be done by hand or with the help of variable filtering techniques.

They implemented 17 classification algorithms, including NB, several variants of DT and other rule base classifiers (OneR, PART, Jrip), LR, and some metaclassifiers, including boosting, bagging, and ensemble schemes.

Furthermore, OneR, the Correlation-based Feature Selection (CFS) approach, the Las Vegas Filter (LVF) algorithm, RELIEF, information gain, and the C4.5 decision tree are all used in this study. To choose the most relevant aspects, the current knowledge domain (from previously published works in BC, medical specialists, and writers' experience) is also studied.

The accuracy, kappa values, AUC, sensitivity, and specificity of the classification findings were all assessed. The algorithms were tested on a small dataset of 257 patients with a high dimensionality of 400 features, which was collected from Iceland's University Hospital (Rose dataset).

The authors went through a feature selection process that yielded three distinct datasets:(1) Base-DS, which contains 98 features based on a medical expert's experience and the outcomes of feature selection techniques; (2) Med-DS, which contains 22 features manually picked from Base-DS by a medical doctor; and (3) Small-DS, which contains just five features hand selected from Base-DS.

The ML findings did not demonstrate to be considerably better than medical professionals' forecasts. The reality that the final sample only comprises categorical variables is discussed in this paper. Even though these data are not given in the study, the authors indicate that their "preliminary analysis" reported no difference variations in prediction outcomes between numerical and discretized versions of some attributes.

The accuracy, sensitivity, specificity, precision, AUC, and Negative Predictive Value (NPV) of the data were assessed using a holdout approach (70 percent –30 percent). In terms of computational models, SVMs and ANNs outperformed the Cox model with the exception of specificity, where SVMs, ANNs, and the Cox model fared similarly: 73 percent, 52 percent, and 94 percent for SVMs, ANNs, and the Cox model, respectively. The best sensitivity (95%) and precision (80%) findings were produced by ANNs, followed by SVMs with 89 percent and 75 percent, respectively. SVMs, on the other hand, proved to be the superior technique, surpassing the others in terms of NPV (89%), accuracy (84.58%), and AUC (84.58%). (0.85). The authors also evaluated the performance of SVMs with the prognostic models St. Gallen's, NPI, and Adjuvan which were previously stated. St. Gallen's had the best sensitivity and NPV (100 percent), but it performed poorly in the remaining tests.

The majority of research works employed local datasets (datasets that solely contain data from a single health center), which makes reproducibility and comparability of results by the other researchers more difficult. Furthermore, the number of patients recruited in most of these trials (less than 1,000) is tiny, particularly for a prevalent disease like BC. Because when majority of the works do not deal with MD at all (and over 80%) or with sufficient completeness, the smaller datasets become quite crucial.

[5] Following extensive talks with numerous medical professionals and literatures, researchers developed a thorough data collecting form with 106 criteria to address the aforesaid problems. Medical specialists supervise the collection of data for the Indian population. Threshold values for various parameters are determined using machine learning algorithms. There are three ranges for each metric that unequivocally reflect the likelihood of becoming diabetes [High, Moderate, or Low]. An Indian Weighted Diabetic Risk Score is determined for each component, such as age, BMI, waist circumference, personal history, family history, food, physical activity, stress, and life quality, based on their proportional risk impact on diabetes. The Total Indian Weighted Diabetes Risk Score has a threshold value that is employed in the diabetic state prediction criterion. There are 106 characteristics and 844 cases in the diabetic dataset utilized in this study.

On the Total Indian Weighted Diabetic Risk Score, density-based clustering using kMean with Euclidean Distance is used. The following are the three clusters that were discovered.

Centroids of clusters: Cluster 1 is at 409.8405, Cluster 2 is at 279.2917, and Cluster 3 is at 344.7333.

StdDev : Cluster 1 is at 20.7733, Cluster 2 is at 27.8148, and Cluster 3 is at 8.2516.

It is found that the lowest cluster centroid is 279.2917.

Randomly selected Test Dataset not used to derive IWDRS with Correct Prediction was 61.5% and Dataset used to derive IWDRS with Correct Prediction was 89.6%.

Randomly selected Test Dataset not used to derive IWDRS Dataset used to derive IWDRS Was 38.5%. Dataset used to derive IWDRS Dataset used to derive IWDRS was 10.4%.

[6] The main goal of this study is to evaluate the performance of some Machine Learning algorithms that are used to predict diabetes diseases. To do so, we use and evaluate four Machine Learning algorithms to predict diabetes mellitus (Decision Tree, K-Nearest Neighbours, Artificial Neural Network, and Deep Neural Network). The dataset was obtained from UCI Repository. There are two portions to these datasets: healthy patients and diabetic patients. The dataset contains 768 occurrences, each with eight attributes/features and one output indicating the patient's label/outcome (0: Not

diabetic, 1: Diabetic).
Diabetes dataset attributes like Pregnancies, Glucose, Blood pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree function, Age, Outcome.
Based on the greatest accuracy, the optimal model for the four classifier methods is chosen. The author employed stochastic gradient descent and feature selection with neighbourhood component analysis (NCA).
To fit the NCA model and plot the feature weights, the best lambda that produces the smallest average loss (0.006) is utilized.
Machine-learning algorithms were used to predict diabetic illnesses in a confusion matrix that was tested on 40 individuals. The Accuracy of ANN 87.5%, DNN is 90%,KNN is 90%,and DT is 82.50.
The Sensitivity of ANN, DNN, KNN and DT are87.71%,100%,91.67%and 71.43% respectively.
The Specificity of ANN, DNN, KNN and DT are 88.46%,84.62%,89.29%and 88.46%.
The research suggest that the DNN achieves a greater accuracy of 90% than the other three Machine Learning algorithms after deleting unnecessary information and lowering dimension.
[7] Based on a selected set of variables from physical examination data, the Author has presented a technique to predict the risk of diabetes. To choose features, he employed Fisher's score, RFE, and a decision tree. The incidence of diabetes was predicted using random forest, logistic regression, SVM, and MLP. Author were able to minimize dimensions with the aid of EA and Fisher's score.
Diabetes has been properly classified using random forest. The findings demonstrate that employing a random forest with 85 characteristics yields the maximum accuracy (0.987). Using Fisher's Score with 19 characteristics, the prediction accuracy was similarly 0.986.
The dataset was gathered from the Physical Examination Center of the Sichuan Provincial People's Hospital. It includes data on 12,848 patients (1,266 diabetics and 11,582 non-diabetics). There are 122 numeric or category type characteristics in each scenario.
Characteristics of Some Variables are Age, gender, BMI, Hypertension, HDL, TG. The classifiers utilized were random forest, logistic regression, SVM, and MLP. The algorithm's performance can be improved by reducing the complexity of the data and using feature selection.
Finally, using the strategy, author has chosen 5 traits to create a new dataset for diabetes prediction. Fasting plasma glucose, HbA1c, HDL, total cholesterol level, and hypertension are the five characteristics. 0.977, 0.968, 0.812, 0.883, 0.875, and 0.905 were the accuracy, precision, sensitivity, F1 score, MCC, and AUC values, respectively.
The results suggest that the strategy may be used to correctly choose features for a diabetes classifier and increase its performance, allowing doctors to detect diabetes more rapidly.
[8] In community settings, the author has identified machine learning (ML) models for type 2 diabetes (T2DM) prediction and determined their prediction accuracy.
Since 2009, the author has done a systematic evaluation of ML predictive modelling experiments in 13 datasets. Discrimination, calibration, and classification measures were among the primary results. Key parameters, amount of validation, and planned usage of models were all secondary results. C-indices meta-analysis, subgroup analyses, meta-regression, publication bias evaluations, and sensitivity analyses were all carried out.
There were 23 investigations (40 prediction models) in all. There were 3, 14, and 6 studies with a high, moderate, or low risk of bias, correspondingly. The models in all of the investigations were internally validated, but none were externally validated. Only seven research attempted model calibration, although twenty studies offered categorization metrics to varied degrees. Eighteen research provided data on both the variables utilized in model creation and the relevance of the features. Twelve studies highlighted the potential for their models to be used in T2DM screening. A excellent pooled c-index was obtained using meta-analysis (0.812). Subgroup analyses and meta-regression were used to identify sources of heterogeneity. There were problems with methodological quality and reporting.
In the community, there is evidence of good performance of ML models for T2DM prediction. Before they're even deployed at scale, they need to enhance their approach, reporting, and validation. There were problems with quality assessment and reporting.
[9] This paper's significant contribution is to give an initial examination on using machine learning algorithms on a publicly available and frequently used diabetes dataset. The Pima Indians Diabetes Dataset was used in this study. The National Institute of Diabetes and Digestive and Kidney Diseases provided the information. The initial purpose of this data was to see if the existence of diabetes could be detected.
The number of pregnancies, glucose level, blood pressure (BP), triceps skinfold thickness (SFT), insulin, BMI, diabetes pedigree function (DPF), and age are all included. There are 768 items in the dataset, 268 of which have hyperglycemia and 500 of which do not. Using the PIMA diabetes dataset, the authors experimented with eight different machine learning algorithms. Neural Networks, SGD, Random Forest, kNN, Nave Bayes, AdaBoost, Decision Tree, and SVM algorithms were used to normalize the data.
First, using stratified 10-fold cross-validation, the procedures were validated. Second, for each approach, the confusion matrix was extracted, and the accuracy, recall, precision, and F1-score were computed. Neural Networks, SGD, and kNN are the three approaches with the best accuracy. The average accuracy between classes for these approaches is 77.47

percent, 76.43 percent, and 73.96 percent, respectively.

[10] The Diabetes Healthcare dataset was taken from the standard benchmark UCI repository and used by the author. There are 9 characteristics and 768 occurrences in the dataset. For diabetes prediction, it was exposed to four different filter-based feature selection approaches. Techniques for filter-based feature ranking have been introduced. To assess classification performance, three distinct classification methods were investigated. RBF Classifier, IBK Classifier, and JRip Classifier are the three distinct Classifiers.

To establish the illness prediction effectiveness, many indicators such as accuracy, TP rate, FP rate, errors (RMSE, MAE, RAE, and RRSE), Kappa Statistics, MCC, and F-Score were analyzed. The findings of this study might help in hyperglycemia prediction and diagnosis in medical science.

[11] The MIMICIII data set was utilised by the author to create and train various models aimed at predicting the death of diabetic patients. Between 2001 and 2012, a repository of de-identified health-related data connected with approximately forty thousand patients who remained in Beth Israel Deaconess Medical Center's critical care units was made publicly available. Tables in the repository include patient admission information, diagnosis, medical notes, and chart events (such as blood glucose levels). The Area Under the Curve (AUC) was the most relevant assessment parameter for these binary classification models, and it was favored over precision in this situation. When comparing to accuracy, which only incorporates the total proportion of true and incorrect predictions, observing the true positive rate (TPR) and false positive rate (FPR) better reveals how the model predicts both occurrences of death. The end findings for each model showed that the CNN (area = 0.885) performed best, followed by the feed forward neural network (area = 0.792) and then random forest (area = 0.771) and decision tree (area = 0.643). These findings support the basic hypothesis that the convolutional network may better comprehend variations in glucose levels. This shows that include blood glucose levels in models can be beneficial. The advances in this article were made possible by the successful application of supervised learning, a convolutional neural network, and enough computer resources.

[12] The article proposes a new enhanced SVM-based ensemble learning model for mellitus detection. The UC Irvine Machine Learning Repository provided the open diabetes dataset (UCL). The University of California, Los Angeles (UCLA) gathered health information from 768 Pima Indian women aged 21 and above. The author evaluates the model using accuracy, recall, and micro-f1 in this work. The benchmark techniques for evaluating our model include classic classification methods (such as SVM, random forest, and C4.5).The classification method K-Nearest Neighbor (KNN). LR is a regression issue for dealing with categorical dependent variables. SVM is a supervised learning model for pattern recognition, classification, and regression analysis.

The mode of the category of the individual tree output determines the category of the random forest's output. Random forest is a classifier that comprises numerous decision trees, and the category of its output is defined by the mode of the category of the individual tree output. C4.5 is a set of algorithms for machine learning and data mining classification. Its purpose is to keep track of what students are learning.

Comparisons with classic SVM experimental result where CLASSIC SVM's recision (mean), recall (mean), Micro-f1 (mean) are 0.6935, 0.5733 and 0.7792 respectively.

The PROPOSED SVM's recision (mean), recall (mean), Micro-f1 (mean) were 0.6893, 0.5743 and 0.7587 respectively.

In the standard SVM training problem, the convex hull technique and the specified limitations reduction strategy are initially utilised to lessen the amount of constraints. Second, the penalty approach is used to turn a constraint-based optimization model into an unconstrained optimization problem.

The best hyperplane is then found by using the gradient descent approach to optimize the objective function. Finally, the improved SVM and additional classification models are combined to produce diagnosis findings.

III. PROPOSED WORK:

Diabetes melitus is important in predicting whether the data leads to cancer or not. Considering the information from [13-Self], the diabetes is predicted from the datasets in various steps. Initially dimensionality reduction is applied on the dataset to remove the redundant data, noisy data or unwanted data. To perform dimensionality reduction, principal component analysis is applied on the features of the dataset. Applying the feature selection and extraction approach fetch the data about the features class and identify the probability to which it belong to. Using the Principal Component Analysis(PCA), the average of the features class is identified to ensure much of the data is not lost. Diabetes prediction can be performed by using the features properties to classify initially whether the dataset is diabetic prone or not. Later after the identification, cancer predictions can be made as per the datasets analysis.

$$R = \frac{P\left(\frac{A > x}{A \cup B}\right) - P\left(\frac{B > y}{A \cup B}\right)}{P(A \cup B)}$$

$$R_x = 1 - \frac{P\left(\frac{A>x}{A\cup B}\right) - P\left(\frac{B>y}{A\cup B}\right)}{P(A\cup B)}$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2758 - 2764
https://publishoa.com
ISSN: 1309-3452

Performing the dimensionality reduction, improves the time and also the computation. The similarity identification is performed using the meta-heuristic algorithm. Identifying similarity is performed by using the fixed threshold of 0.64. Standard deviation for the computation is fixed initially with 0.5. Clustering is applied by using the similarity measure to form the common groups. Clustering is re-computed on the dataset, until the final clusters are formed. Once the initial clusters and final clusters found to be similar, the computation is stopped and final mean and standard deviation is calculated for the clusters group. This iterations are performed to identify better accuracy and precision.

IV. RESULTS

Using the PIMA, UCI and AIM-94 dataset, the computation is performed. Dimensionality reduction and similarity identification is performed using various pre-existing measures to compute the said datasets and found the various accuracies, precisions and recalls for the given dataset.

The computation is performed using Naïve Bayes, Random Forest, Random Tree and proposed measure. The proposed measure shows the better results compared to the existing similarity measure by computing the clustering and classification.

Table 1. Confusion Matrix of Naïve Bayes with Aim-94 dataset

Classification	Confusion Matrix				Metric						
		Class 1	Class 0	Total	TP	FN	FP	TN	Accuracy	Precision	Recall
Naïve Bayes	Class 1	1325	980	2305	0.57	0.43	0.14	0.86	0.67	0.89	0.57
	Class 0	167	1006	1173	0.14	0.86	0.00	0.00	0.14	1.00	0.14
				3478	2264	1214	1214	2264			

Table 2. Confusion Matrix of Random Tree with Aim-94 dataset

Classification	Confusion Matrix				Metric						
		Class 1	Class 0	Total	TP	FN	FP	TN	Accuracy	Precision	Recall
Random Tree	Class 1	1692	510	2202	0.77	0.23	0.40	0.60	0.71	0.77	0.77
	Class 0	515	761	1276	0.40	0.60	0.00	0.00	0.40	1.00	0.40
				3478	2453	1025	1025	2453			

Table 3. Confusion Matrix of Random Forest with Aim-94 dataset

Classification	Confusion Matrix				Metric						
		Class 1	Class 0	Total	TP	FN	FP	TN	Accuracy	Precision	Recall
Random Forest	Class 1	1762	428	2202	0.80	0.20	0.34	0.66	0.75	0.80	0.80
	Class 0	440	848	1276	0.34	0.66	0.00	0.00	0.34	1.00	0.34
				3478	2610	868	868	2610			

Table 4. Confusion Matrix of Proposed Approach with Aim-94 dataset

Classification	Confusion Matrix				Metric						
		Class 1	Class 0	Total	TP	FN	FP	TN	Accuracy	Precision	Recall
Proposed Approach	Class 1	2369	130	2499	0.95	0.05	0.43	0.57	0.84	0.85	0.95
	Class 0	419	560	979	0.43	0.57	0.75	0.66	0.43	1.00	0.43
				3478	2929	549	549	2929			

From the above said results, it is observed that, the proposed approach shows better accuracy of 84% compared to existing approaches and with this the method gives scope to work on the prediction of cancer from the dataset. The Naïve bayes, Random Forest, Random Tree gives less accuracy in predicting the diabetes and cancer.

V. CONCLUSION

The approach towards identification of diabetes and cancer has predominantly grown and found many approaches have already shown the prediction results. The proposed approach is computed on AIM-94 dataset to identify the similarity of the dataset. Using Meta-Heuristic approach, the features are identified and similarity is identified based on the class of the record. Computation is performed to apply dimensionality reduction, classification and clustering techniques and found the proposed approach produces better accuracy compared to the existing methods.

REFERENCES

1. Xia, Chao, et al. "A convolutional neural network based ensemble method for cancer prediction using DNA methylation data." Proceedings of the 2019 11th International Conference on Machine Learning and Computing. 2019.

2. Xiao, Yawen, Jun Wu, Zongli Lin, and Xiaodong Zhao. "A deep learning-based multi-model ensemble method for cancer prediction." Computer methods and programs in biomedicine 153 (2018): 1-9.

3. Zafar, Faizan, et al. "Predictive analytics in healthcare for diabetes prediction." Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology. 2019.

4. Abreu, Pedro Henriques, et al. "Predicting breast cancer recurrence using machine learning techniques: a systematic review." ACM Computing Surveys (CSUR) 49.3 (2016): 1-40

5. Omprakash Chandrakar and Jatinderkumar R. Saini. 2016. Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes. In Proceedings of the 9th Annual ACM India Conference (COMPUTE '16). Association for Computing Machinery, New York, NY, USA, 125–128. DOI:https://doi.org/10.1145/2998476.2998497.

6. Daanouni, Othmane, Bouchaib Cherradi, and Amal Tmiri. "Diabetes Diseases Prediction Using Supervised Machine Learning and Neighbourhood Components Analysis." Proceedings of the 3rd International Conference on Networking, Information Systems & Security. 2020.

7. Jiaqi Hou, Yongsheng Sang, Yuping Liu, and Li Lu. 2020. Feature Selection and Prediction Model for Type 2 Diabetes in the Chinese Population with Machine Learning. Association for Computing Machinery, New York, NY, USA, Article 103, 1–7. DOI:https://doi.org/10.1145/3424978.3425085.

8. Silva K, Lee WK, Forbes A, Demmer RT, Barton C, Enticott J. Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. Int J Med Inform. 2020 Nov;143:104268. doi: 10.1016/j.ijmedinf.2020.104268. Epub 2020 Sep 7. PMID: 32950874.

9. Marques, Gonçalo, Ivan Miguel Pires, and Nuno M. Garcia. "Diabetes Disease through Machine Learning: A comparative study." 2020 4th International Conference on Computer Science and Artificial Intelligence. 2020.

10. Mishra, Sushruta & Chaudhury, Pamela & Mishra, Brojo & Tripathy, Hrudaya. (2016). An implementation of Feature ranking using Machine learning techniques for Diabetes disease prediction. 1-3. 10.1145/2905055.2905100.

11. Ian Wittler, Xinlian Liu, and Aijuan Dong. 2019. Deep Learning Enabled Predicting Modeling of Mortality of Diabetes Mellitus Patients.Association for Computing Machinery, New York, NY, USA, Article 100, 1–6. DOI:https://doi.org/10.1145/3332186.3333262

12. Yang, Zihe, Yinghua Zhou, and Chenxu Gong. "Diagnosis of diabetes based on improved Support Vector Machine and Ensemble Learning." Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence. 2019.