Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.755183

DIABETES DIAGNOSIS USING MACHINE LEARNING AND DATA VISUALIZATION

Shivam Singh
School of Computer science and engineering, Lovely Professional University, Phagwara, India, shivamsingh45978@gmail.com

Nahita Pathania
School of Computer science and engineering, Lovely Professional University)
Phagwara, India, nahita.19372@lpu.co.in

Novel Biswas
School of Computer science and engineering, Lovely Professional University
Phagwara, India, novelbiswas@gmail.com

S.Guru Vishnu Vardhan Raju
School of Computer science and engineering, Lovely Professional University
Phagwara, India, vishnuraju7997@gmail.com

Dipti Gaurav Mishra
School of Computer science and engineering, Lovely Professional University
Phagwara, India, diptigaurav1015@gmail.com

Abstract—The burgeoning prevalence of diabetes globally necessitates swift diagnosis and effective intervention for public health benefit. Leveraging data from the Behavioural Risk Factor Surveillance System (BRFSS) dataset, the current piece of research is trying to figure out if machine learning algorithms can be used for diabetes detection. By analysing a range of health indicators including blood pressure, cholesterol, BMI, lifestyle factors, and socio-demographic features, the research employs algorithms for example Logistic Regression, CatBoost, K-Nearest Neighbours (KNN), Random Forest, Decision Tree to predict diabetes status. Through meticulous data preparation involving dealing with absent data, transforming categorical variables into numerical representations, and standardizing numerical attributes, followed by evaluation using designated test subsets, the study aims to delineate the most efficient algorithmic strategies for diabetes diagnosis, contributing insights to healthcare analytics and proactive disease management.

Keywords—Machine Learning, Diabetes diagnosis, Health indicators, Cat Boost, Random Forest, F1-score, Evaluation.

I. INTRODUCTION

Diabetes mellitus, an enduring metabolic condition marked by elevated levels of blood sugar, represents a considerable global health issue. Its incidence continues to rise, underscoring the urgent need for timely diagnosis and effective treatment approaches. Diabetes exists in various forms, with type 1 diabetes stemming from inadequate insulin production and type 2 diabetes linked to insulin resistance and reduced insulin effectiveness. Without proper intervention, diabetes can trigger numerous Complications such as cardiovascular issues, renal concerns, and nerve deterioration contribute to the condition severe health complications and increased mortality rates.

With the escalating challenges posed by diabetes, there's been a focused endeavor to seek novel methods to enhance its detection and treatment. Owing to the great leap in the application of machine learning in the medical field, utilizing the tools which are provided by machine learning is a more promising and emerging path in improving the diagnostic accuracy and prediction across many medical fields. into extensive collections of Health-related information like Behavioral Risk Factor Surveillance System (BRFSS) dataset, machine learning [5] algorithms present a powerful resource for extracting valuable insights and assisting healthcare professionals in making informed decisions.

The BRFSS dataset, a comprehensive repository of health indicators, encompasses a wide range of variables pertinent to diabetes assessment. These include parameters like blood pressure, cholesterol levels, body mass index (BMI), lifestyle elements, and sociodemographic characteristics. This study aims to leverage machine learning capabilities to utilize the abundant data available and create predictive models for diagnosing diabetes. Employing a diverse array of machine learning practitioners utilize techniques such as Decision Trees, Random Forest and Logistic Regression. Cat Boost, and K-Nearest Neighbors (KNN),[1] seek to leverage these health indicators to accurately predict diabetes status.

Before training and assessing the models, thorough preprocessing steps are carried out to ensure the accuracy and consistency of the data. The missing data is managed, binary categorical variables are converted in to the format suitable for the models, and the numerical features are standardized for the best performance of the models, too [11]. Using systematic search machine learning algorithms makes the machine learning more efficient and effective techniques, this study aims to offer valuable insights into their potential for diagnosing diabetes. Equipping healthcare professionals with reliable predictive tools can aid in early detection and intervention, ultimately enhancing patient outcomes and advancing diabetes management. However, while promising, the theoretical application of current models in practical healthcare circumstances of the real world is hindered by the fact that these models need more validation and refinement through dedicated research efforts.

II. LITERATURE REVIEW

[1] Farajollahi et al. devised a machine learning-driven strategy for diabetes diagnosis, achieving 83% accuracy with Adaboost. They compared various classifiers including Decision Tree, logistic regression, Support Vector Machine (SVM), Adaboost and Random Forest alongside latter demonstrating the highest accuracy.

[2] Kalia et al. designed a web-based diabetes prediction system employing methods such as Naïve Bayes, Logistic Regression, K-nearest neighbors, Random Forest, Support Vector Classifier and Decision Tree. They utilized preprocessing techniques and data visualization to boost the accuracy of classification.

[3] Derozier et al. proposed a method using qualitative color scales and machine learning algorithms to display glycemic data effectively for diabetes management. They demonstrated the utility of their approach in highlighting important glycemic patterns.

[4] Cho et al. applied logistic regression and SVM for predicting diabetic nephropathy, achieving high accuracy with linear SVM classifiers. They employed feature selection methods to improve classification performance.

[5] Arumugam et al. investigated multi-disease indicator using various algorithms that machine learning use, with decision trees consistently outperforming other models. Their study emphasized the significance of data mining in healthcare for predicting diseases.

[6] Kavakiotis et al. Conducted a comprehensive examination of applications that machine learning use in the context of diabetes. research, with support vector machines being the most successful algorithm. They highlighted the usefulness of extracting valuable knowledge from clinical datasets.

[7] Ak et al. they conducted a relative assessment of breast cancer identification, attaining the top prediction precision of 98.1%. with logistic regression. They emphasized the importance of accurate diagnosis in breast cancer treatment.

[8] Mujumdar and Vaidehi proposed a diabetes prediction model incorporating external factors, demonstrating enhanced classification accuracy compared to existing methods. They highlighted the role of Leveraging big data analytics to enhance medical care outcomes.

[9] Warke et al. explored the significance of data mining in the medical care sector and put forward a model for predicting diabetes that showed improved classification accuracy. They emphasized the potential of big data analytics in transforming medical care sector.

[10] McCarthy et al. explored application of visualization techniques in management, diagnosis, and cancer detection. Their study emphasized significance exploratory data analysis techniques in extracting clinically useful knowledge from biological data.

[11] Chen et al. explored the design of visualizations to aid non-expert machine learning practitioners in diagnosing model problems. Their study focused on improving interpretability and explainability in the decision-making process of machine learning models.

[12] Bruno et al. designed a system for understanding automatic diagnosis processes using visual analysis techniques. Their framework aimed to aid in debugging, interpreting, and contrasting machine learning models transparently and interactively.

[13] Zhang et al. introduced Diverse, a structure independent of specific models, designed for diagnosing and interpreting diagnosing machine learning models. Their approach employed visual analysis techniques to assist in comparing machine learning models , debugging and interpreting.

[14] Zhou et al integrated data visualization with support vector data description and the glowworm swarm optimization algorithm. for early detection of liver disease. Their approach achieved high sensitivity, specificity, and accuracy in diagnosing early liver disease.

The literature review reveals that machine learning techniques are vital in diagnosing and managing diseases across different medical fields. Studies have demonstrated the effectiveness of machine learning algorithms such as logistic regression, support vector machines, random forests, Adaboost and decision trees in accurately diagnosing diseases including liver diseases, diabetes and breast cancer. Additionally, the integration of data visualization techniques has proven to be instrumental in enhancing the interpretability and explainability of machine learning models, thereby facilitating diagnosis and decision-making processes. Furthermore, the development of model-agnostic frameworks like Manifold offers promising avenues for interpreting and diagnosing machine learning models across different domains. Overall, these The results highlight the significance of utilizing data visualization techniques and machine learning in healthcare to improve diagnostic accuracy, enhance patient outcomes, and ultimately advance medical research and practice.

III. METHODOLOGY

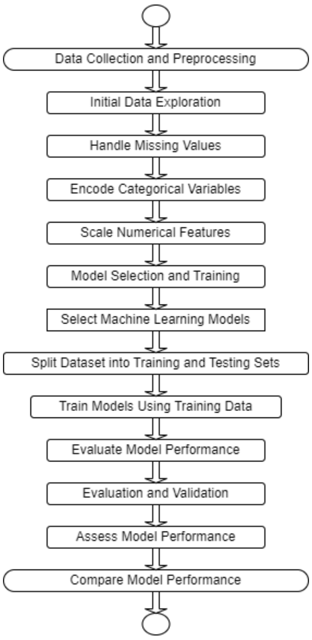
Data Collection and Preprocessing
Initial Data Exploration
Handle Missing Values
Encode Categorical Variables
Scale Numerical Features
Model Selection and Training
Select Machine Learning Models
Split Dataset into Training and Testing Sets
Train Models Using Training Data
Evaluate Model Performance
Evaluation and Validation
Assess Model Performance
Compare Model Performance

Fig. 1. Proposed Methodology flowchart.

The flow chart can be explained in the subsequent context;

A. Preprocessing and Data Collection

The initial step involves collecting the dataset and performing preprocessing tasks to ensure data integrity. This involves managing absent data, converting categorical variables into numerical ones, and standardizing numerical characteristics. In given data set, found 23899 duplicates values which were dropped. These preparatory steps are crucial for the subsequent analysis. As such duplicate values can change the output.

B. Checking for the outliers

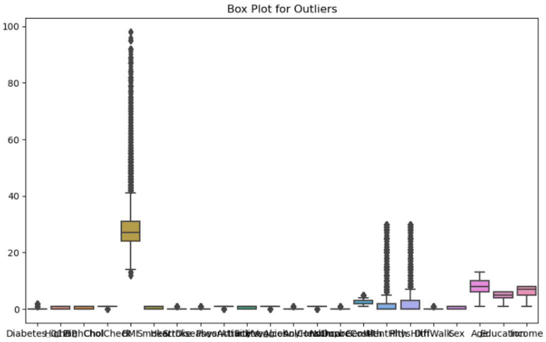

Fig. 2. Box plot for outliers.

There are outliers in BMI and PhysHlth but cannot remove all of them as they are in large quantity so need to fix z range for the box plot to reduce the outliers. Let's explore the data first.

C. Initial Data Exploration

Following preprocessing, an exploratory analysis of the dataset is conducted to gain insights into its structure and characteristics. Descriptive statistics, visualizations, and other exploratory techniques are employed to understand the distribution and relationships within the data.

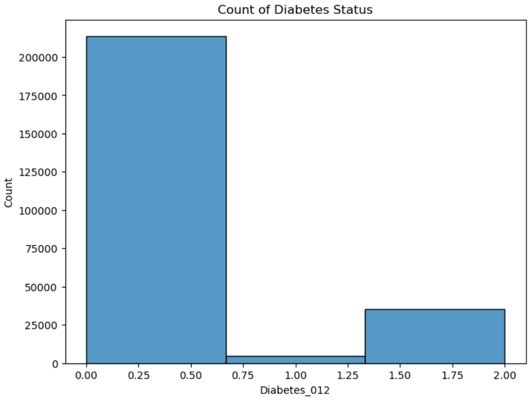

Fig. 3. Count of Diabetes Status

Fig. 3. is a plot of count of people along with the diabetes range. So, have the range from 0 to 2. By looking at figure can see that divide this in to 3 ranges.

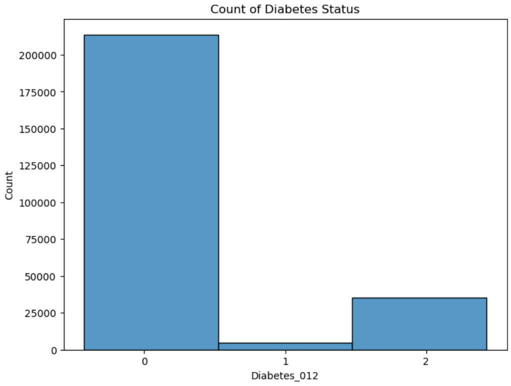

Fig. 4. Count of Diabetes Status

In Fig. 4. divided the data in to 3 classes, that is the range which is 0,1, and 2.

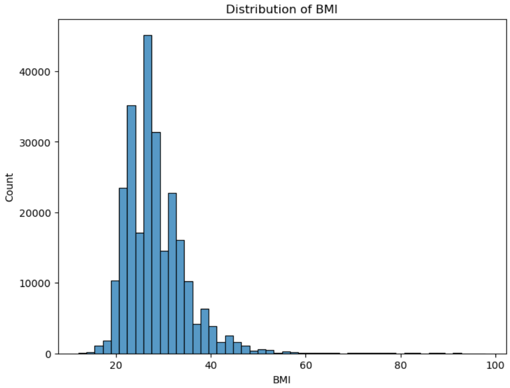

Fig. 5. Distribution of BMI

From Fig. 5. from BMI score 20-40 the count of the people is higher.

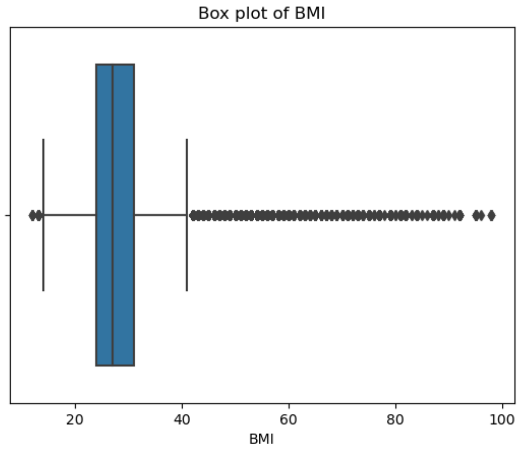

Fig. 6. Box plot for BMI

BMI plays an important role that need to check if BMI contains outliers, that is the reason why plotted Box plot in Fig. 6. After seeing these many numbers of outliers adjusted the upper and lower limits.

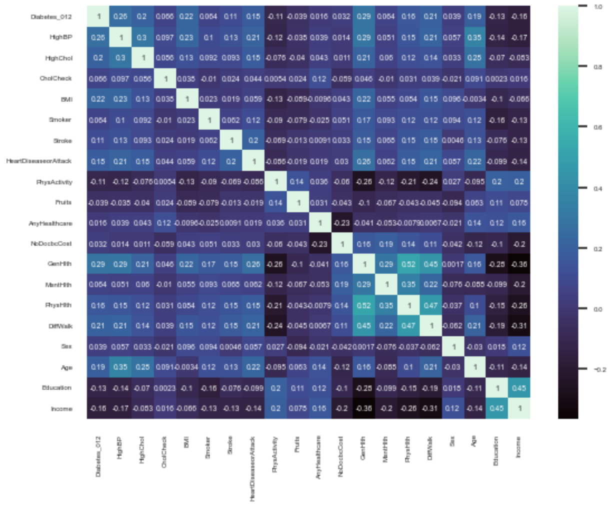

Fig. 7. Confusion Matrix

BMI value importance, so plot Confusion Matrix in Fig. 7. to see the relation between each element.

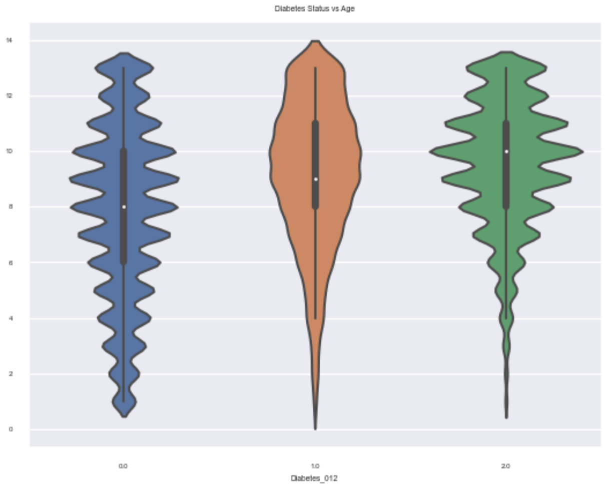

Fig. 8. Diabetes Status Vs Age.

From Fig. 8. Diabetes status 1 the Age group is stable when compared to Age groups for 0 and 2. For status 0 the Age group is most unstable when compared to 1 and 2.

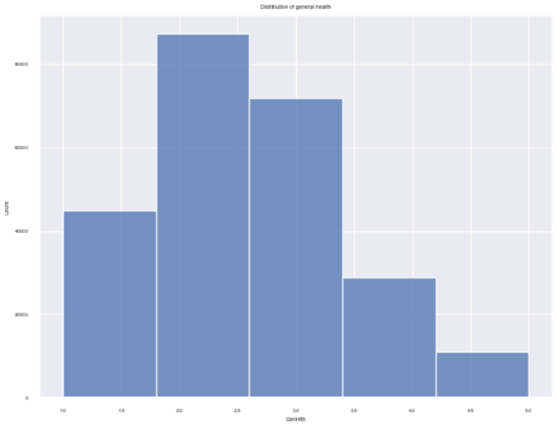

Fig. 9. Distribution of general health.

Fig. 9. Is a plot of GenHealth and count, found that value from 2 to 2.5 has the highest count in the dataset.

D. Model Selection and Training

Machine learning models are selected based on the nature of the problem and the characteristics of the dataset. Various algorithms such as Random Forest, K-Nearest Neighbors (KNN), Cat Boost, Logistic Regression and Decision Tree are considered for training.

TABLE I. ACCURACY TABLE

Index	Model Name	Accuracy
1.	KNN	0.8330
2.	Cat Boost	0.8501
3.	Random Forest	0.844
4.	Logistic Regression	0.847
5.	Decision Tree	0.7704

Cat Boost performs better than any other model as it has gained an accuracy rate of 85.01%. Other than Cat Boost Logistic Regression and Random Forest has the second highest correctness ratio that is 84.1% and 84.7% respectively.

E. Split Dataset into Testing Sets and Training sets

It is divided into a pair parts: one for training the models and another for testing their performance. It is done such that the model is shown part of the data to be taught and then it has to be tested with data that has not been detected to check whether the model has the generalization capability.

F. Train Models using Training Data

This refers to the chosen machine learning models that are trained with the training data. The models make such adjustments by wherein they can be able to recognize structures and the relations that are captured in the data.

G. Evaluate Model Performance

Then, the performance of each model is measured against different metrics specified by F1-score, precision, accuracy and recall. This process aids assess how well the models are able to predict diabetes status based on the provided health indicators.

TABLE II. MODEL REPORT

Model	F1-score	Precision	Recall	Accuracy
Decision Tree	0.77	0.78	0.77	0.77
Logistic Regression	0.80	0.80	0.84	0.84
Random Forest	0.80	0.79	0.84	0.84
Cat Boost	0.81	0.80	0.85	0.85
KNN	0.80	0.78	0.83	0.83

Upon comparing the evaluation metrics for each model, it's evident that CatBoost outperforms the other models with regard to F1 score, recall, precision, and accuracy. With F1-score of 81.33%, recall of 85.01%, precision of 80.56%, and an accuracy of 85.01%, CatBoost demonstrates superior predictive performance compared to the other models. Logistic Regression also exhibits commendable results with an F1-score of 80.98%. and an accuracy of 84.79% Random Forest follows closely behind with and an F1-score of 80.85% and an accuracy of 84.13%. However, Decision Tree shows Marginally reduced performance with an F1-score of 77.51% and an accuracy of 77.04%. K-Nearest Neighbors (KNN) presents moderate results with an F1-score of 80.44% and an accuracy of 83.30% Overall, CatBoost emerges as the top-performing model in this analysis, offering the highest F1 and accuracy performance level compared to other models evaluated.

ABOUT DATA SET

A. Diabetes Health Indicators

The dataset employed in this project comprises health indicators extracted from the Behavioral Risk dataset. It includes various metrics related to diabetes and general health, encompassing factors including blood pressure, cholesterol levels, smoking habits, body mass index (BMI), physical activity, and socio-demographic characteristics.

B. Dataset Details:

a) Name: Diabetes Health Indicators (BRFSS dataset)
b) Number of Rows: 253,680
c) Source: Behavioral Risk Factor Surveillance System
d) Number of Columns: 22

C. Key Features:

a) Diabetes_012: Indicates diabetes status (target variable).
b) High BP: Presence of high blood pressure.
c) High cholesterol: Presence of high cholesterol levels.
d) CholCheck: Whether cholesterol levels were checked.
e) BMI: Body mass index.
f) Smoker: Smoking habits.
g) Stroke: History of stroke.
h) HeartDiseaseorAttack: History of heart disease or heart attack.
i) PhysActivity: Engagement in physical activity.
j) Fruits: Consumption of fruits.
k) Veggies: Consumption of vegetables.
l) HvyAlcoholConsump: Heavy alcohol consumption.
m) AnyHealthcare: Access to healthcare services.
n) NoDocbcCost: Healthcare coverage without out-of-pocket costs.
o) GenHlth: General health status.
p) MentHlth: Mental health status.
q) PhysHlth: Physical health status.
r) DiffWalk: Difficulty in walking.
s) Sex: Gender.
t) Age: Age of respondents.
u) Education: Educational level.
v) Income: Household income.

IV. CONCLUSION

In conclusion, this study particularly deals with machine learning algorithms and their use in diagnosing diabetes among respondents of the BRFSS survey. began by acknowledging the growing prevalence of diabetes as a global health concern, emphasizing the need for timely diagnosis and effective management strategies. Leveraging machine learning techniques, including Logistic Regression, Random Forest, CatBoost, K-Nearest Neighbors (KNN), Decision Tree predict diabetes status based on a diverse array of health indicators encompassed in the BRFSS dataset.

Prior to model training and evaluation, rigorous preprocessing steps were undertaken to ensure the integrity and reliability of the data. This included scaling numerical features, addressing missing values, encoding categorical variables and addressing missing values to optimize model performance. Through systematic exploration of machine learning methodologies, aimed to provide valuable insights into the potential of these algorithms in diabetes diagnosis.

As moving forward, there are several avenues for further development and enhancement of this project. One key direction involves the integration of additional data sources to enrich the predictive capabilities of the models. Data stemming from wearable devices, electronic health records and genetic information can be pooled to provide a more holistic and thorough health agenda and hence enhance the precision of diabetes diagnosis.

Additionally, the deployment of these models in real-world clinical settings presents an exciting opportunity for validation and refinement. By working in partnership with health providers and in assisting them to carry out prospective research, we can certainly learn more about the practical efficiency of these models in supporting clinical decisions.

Moreover, the development of user-friendly interfaces and decision support systems can facilitate seamless integration of these predictive tools into routine clinical practice. By streamlining the implementation process and enhancing user accessibility, maximize the impact of these models in improving patient care and outcomes.

Overall, the future of this project lies in continued innovation, collaboration, and validation efforts, with the ultimate goal of harnessing the full potential of machine learning in revolutionizing diabetes diagnosis and management.

Our results demonstrated promising outcomes, with CatBoost achieving the highest an F1-score of 81.33% and accuracy of 85.01% outperforming other models in terms of predictive performance. Logistic Regression also demonstrated strong performance, achieving an F1-score of 80.98% and an accuracy of 84.79%. These findings underscored the capability of machine learning methods to support medical care experts in their work. diagnosing diabetes, thereby facilitating early intervention and improved patient outcomes.

REFERENCES

[1] Boshra Farajollahi, Maysam Mehmannavaz, Hafez Mehrjoo, Fateme Moghbeli, Mohammad Javad Sayadi, Diabetes Diagnosis Using Machine Learning, DOI: https://doi.org/10.30699/fhi.v10i1.267, ISSN-Online: 2676-7104, Vol10(2021).

[2] A. R. Kalia, A. Pavshe, D. Shah and S. Pansambal, "Data visualization and pre-processing techniques based Diabetes Prediction System," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 1638-1645, doi: 10.1109/ICESC51422.2021.9532964.

[3] Derozier, V., Arnavielhe, S., Renard, E., Dray, G., & Martin, S. (2019). How Knowledge Emerges From Artificial Intelligence Algorithm and Data Visualization for Diabetes Management. Journal of Diabetes Science and Technology. https://doi.org/10.1177/1932296819847739

[4] Cho, B. H., Yu, H., Kim, K., Kim, T. H., Kim, I. Y., & Kim, S. I. (2007). Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artificial Intelligence in Medicine, 42(1), 37-53. https://doi.org/10.1016/j.artmed.2007.09.005

[5] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2022). Multiple disease prediction using Machine learning algorithms. Materials Today: Proceedings, 80, 3682-3685. https://doi.org/10.1016/j.matpr.2021.07.361

[6] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2016). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005

[7] Ak, M. F. (2020). A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare, 8(2), 111. https://doi.org/10.3390/healthcare8020111

[8] Mujumdar, A., & Vaidehi, V. (2018). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science, 165, 292-299. https://doi.org/10.1016/j.procs.2020.01.047

[9] Mitesh Warke, Vikalp Kumar, Swapnil Tarale, Payal Galgat, D.J Chaudhari, Diabetes Diagnosis using Machine Learning Algorithms e-ISSN: 2395-0056 Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

[10] MCCARTHY, J. F., MARX, K. A., HOFFMAN, P. E., GEE, A. G., UJWAL, M. L., & HOTCHKISS, J. (2004). Applications of Machine Learning and High-Dimensional Visualization in Cancer Detection, Diagnosis, and Management. Annals of the New York Academy of Sciences, 1020(1), 239-262. https://doi.org/10.1196/annals.1310.020

[11] D. Chen, R. K. E. Bellamy, P. K. Malkin and T. Erickson, "Diagnostic visualization for non-expert machine learning practitioners: A design study," 2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), Cambridge, UK, 2016, pp. 87-95, doi: 10.1109/VLHCC.2016.7739669.

[12] P. Bruno, F. Calimeri, A. S. Kitanidis and E. D. Momi, "Understanding Automatic Diagnosis and Classification Processes with Data Visualization," 2020 IEEE International Conference on Human-Machine Systems (ICHMS), Rome, Italy, 2020, pp. 1-6, doi: 10.1109/ICHMS49158.2020.9209499.

[13] J. Zhang, Y. Wang, P. Molino, L. Li and D. S. Ebert, "Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models," in IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 1, pp. 364-373, Jan. 2019, doi: 10.1109/TVCG.2018.2864499.

[14] Zhou, X., Zhang, Y., Shi, M., Shi, H., & Zheng, Z. (2014). Early detection of liver disease using data visualization and classification method. Biomedical Signal Processing and Control, 11, 27-35. https://doi.org/10.1016/j.bspc.2014.02.006