j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 8, issue 4. 2025

A Comprehensive Data Enhancement Method for the Pima Dataset to Improve Diabetes Prediction Performance

Mohammed F. Zamil[*1], Dunia H. Hameed[2], Usama Samir Mahmoud[3]

[1&2] *Biomedical Information College, University of Information Technology and Communications, Baghdad, Iraq.*

[3]*University of Information Technology and Communications, Baghdad, Iraq..*

mfadhil@uoitc.edu.iq

Abstract Diabetes is one of the silent killer diseases that can effect if left without medication and a real change in lifestyle. 10.5% of adult people (10-79 years) have diabetic in the world according to the International Diabetes Federation (IDF) Diabetes Atlas (2021) reports [1]. And number getting higher. Thus, in this study, we aim to build a prediction model using Pima Indian Diabetes (PID) dataset. Dataset required heavy-duty processing because of its low-quality characteristics, such as lot missing values and imbalance. This paper shows how enhancing data quality can affectively reflect on models' performance. Based on the conducted experiments, ensemble models such as Random Forest show highest performance (0.86% AUC-ROC) with highest encasement among all other model by around 4%.

Crossref 10.36371/port.2025.4.1

Keywords: *Diabetes Prediction, Pima Indian Diabetes (PID) Dataset, ensemble Models.*

1. INTRODUCTION

Diabetes is a chronic disease that can be diagnosed by elevated blood glucose levels. It has become one of the most challenging global health issues. According to statistics presented by the International Diabetes Federation (IDF), globally there will be around 537 million adults living with diabetes in 2021, and statistics predict that this number will rise to 783 million by 2045 [1]. The disease has a negative impact on both patients and healthcare systems, as this disease was a primary or contributing factor to the mortality of an estimated 6.7 million deaths. Beyond its health implications, the disease also posed economic challenges for healthcare organizations, as around 20% of money spent either directly goes to diabetes or diabetes-related diseases [2]. To avoid the aforementioned negative effects, early detection is crucial to prevent or at least get control of its consequences, such as retinopathy, heart disease, and kidney disorder [3].

Diagnosing diabetes with high confidence is quite a challenge and requires laboratory testing alongside clinical symptom assessment. Despite all these tests and patient-reported symptoms, repeated laboratory tests may be required for confirmation. So, these excessive costs may limit access to proper diagnosis. The World Health Organization (WHO) statistics show that almost half of diabetic people remain undiagnosed [4]. Thus, there is a crucial need for more easy-to-access, less time-consuming and cost diagnosing tools.

Machine learning (ML) has become one of the essential tools in the healthcare industry and has enabled the development of disease prediction and diagnosis. ML abilities to identify complex patterns in medical data utilized for tasks such as enhanced disease diagnosis [5], personalized treatment [6], and predicting patient outcomes (i.e., disease progression, survival rate, and recovery time) [7]. Optimizing clinical decision-making by machine learning leads to overall healthcare services enhancement [8]. Despite using machine learning in detecting diabetes showing lot possibilities, there is lack of high-quality real-world datasets. As existing models for diabetes detection suffer from limitations such as low generalization power as it trained on small, imbalanced, and biased datasets.

In this study we aim to build diabetes detection models using Pima Indian Diabetes (PID) dataset [9]. Despite its popularity in the academic researcher's community, the dataset has limitations such as small size, imbalance, gender, and race biased. Therefore, this study aims to employ intensive preprocessing techniques to enhance quality or at least mitigate data quality concerns. Then six light weight machine learning models applied such as Random Forest, logistic regression, and Artificial Neural Networks. And to evaluate models we used metrics such as accuracy, precision, recall, and F1-score. By addressing data limitations, this research seeks to contribute to the development of generalizable and high-performance models.

2. RELATED WORK

Early diabetes detection is one of the main goals to enhance health quality, Artificial Intelligence (AI) and Machine Learning (ML) become important tools for researchers to study datasets and build models. Many previous studies have used machine learning to extract patterns from patient data to predict diabetes risk. Studies showed that ML methods can effectively identify patient data with diabetics. Most studies explore new ML algorithms and datasets. However, having a reliable and good enough dataset remains the main challenge in building a reliable and generalized model. This section presents the previous efforts for the early stage of diabetes disease prediction.

Abnoosian K., et al. [10] designed a predictive diabetes framework based on classification to three classes: diabetic, non-diabetic, and prediabetes. They utilized Iraqi Patient Dataset for Diabetes (IPDD) which is imbalanced data set, so several pre-processing methods implemented and they merged various machine learning models (such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), AdaBoost, and Gaussian Naive Bayes (GNB)) to manipulate the misbalancing problem. Diabetes. The resulted model had high average accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC) values of 0.9887, 0.9861, 0.9792, 0.9851, and 0.999, respectively.

Viswanatha V., et al.[11] proposed a predictive diabetes model to know if the person will have diabetes or not based on specific measures in the datasets. They implemented Logistic regression algorithm and used Python IDEs for analysis. Feature Selection methods were used in addition to aggregation method Maximum Voting. They used two datasets: the first dataset from Pima Indians Diabetes dataset for implementing aggregation Maximum Voting method which showed about 78% accuracy. The second dataset from Vanderbilt used incorporated techniques showed about 93% accuracy.

Saihood Q., et al. [12] introduced framework for diabetes prediction contained three stages. These stages include preprocessing, Machine Learning usage for hyperparameters which tuned correctly by implementing grid search and the last stage carried out ensemble techniques such as bagging, boosting and stacking to improve the productivity. They used Pima dataset for testing the framework. The accuracy of stacking method with the stacked Random Forest (RF) and Support Vector Machine (SVM) model resulted 97.50% while the accuracy of the bagging methods with the RF model resulted 97.20%. For the boosting methods, eXtreme Gradient Boosting (XGB) model achieved the accuracy was 97.10%.

Ganie S., et al. [13] executed five boosting algorithms on Pima dataset. Several methods were used for predictive analytics: up sampling, normalization, feature selection, and hyperparameter tuning. The results analysis implemented by using statistical/machine learning metrics and k-fold cross-validation techniques. Gradient boosting resulted accuracy rate of 92.85%. Precision, recall, f1-score, and Receiver Operating Characteristic (ROC) curves were used for testing.

Reza M., et al. [14] introduced two stacking ensemble models by using neural networks with two types classical and deep type, they used collection of data obtained from Pima dataset, simulated data and a local healthcare data. The results presented high accuracy for early detection of diabetes.

Kalyani K., et al. [15] used data from Kaggle to diagnose early diabetes and predict if the person will develop diabetes or not. They implemented a random forest algorithm for prediction analysis. They found that this algorithm was reliable and effective.

Khan Q., et al. [16] proposed classifiers such as Support Vector Machine (SVM), Neural Networks, a Gradient Boosting (GB), Tab-Net and a Random Forest (RF) algorithms, they used three methods for choosing the features from Pima and early-risk datasets. They showed high accuracy results.

Salih M., et al [17] designed a model by using several methods like Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT). They used Pima dataset with preprocessing consisted from two stages. Principal Component analysis (PCA) was used as feature selection method. This integration exhibited high accuracy.

From the presentation of the above researches, we conclude that the pre-processing for the dataset was very useful and the stacking ensemble machine learning model exhibited promising results.

3. DESCRIPTION AND CHALLENGES OF THE DATASET

The dataset used in this study is the Pima Indian Diabetes (PID) dataset, it is publicly available on the UCI Machine Learning Repository [9]. Data collection conducted by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The study targeted the female Pima Indian population, a community with a high prevalence of diabetes [18]. Although the study that collected the dataset was published in 1988, Pima still relevant in the current research community working on diabetes prediction. A query of Google Scholar identified 2230 studies in 2024 that mentioned the ‘PIMA dataset’, reflecting its high presence in recent scientific publication.

Pima is a tabular dataset of medical health information for female patients of Pima Indian heritage from 21 years old and older. The dataset contains 768 rows and 9 columns. Each row represents a female patient been record their data such as glucose levels, blood pressure, BMI, and age, bellow table (1) summaries dataset columns—the dataset provides a foundation for developing predictive models to identify individuals at risk of diabetes. Despite its utility, the dataset presents several challenges, including missing values represented as

M. F. Zamil, et al. 2025, A Comprehensive Data Enhancement Method for the Pima Dataset to Improve Diabetes Prediction Performance. *Journal port Science Research*, 8(4), pp.314-320. https://doi.org/10.36371/port.2025.4.1

implausible zeros, class imbalance between diabetic and non-diabetic cases, and a relatively small sample size that limits scalability.

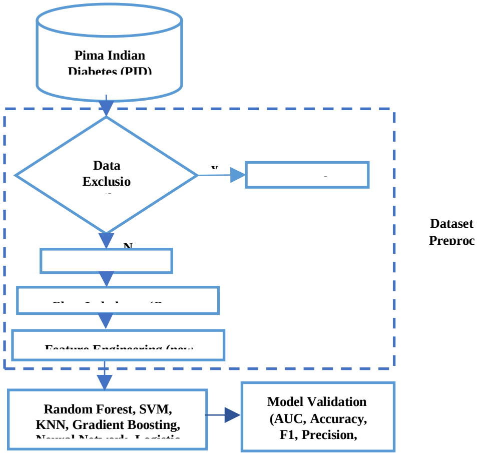
Table 1: *Pima features dataset description*

Features	*Notes*	*Count*	*Median*	*Mean*	*Std*	*Min*	*Max*
Pregnancies	No missing values	768	3	3.845	3.3695	0	17
Glucose	Anormal (5) zero values	768	117	120.895	31.972	**0**	199
Blood Pressure	Anormal (35) zero values	768	72	69.105	19.355	**0**	122
Skin Thickness	Anormal (227) zero values	768	23	20.536	15.952	**0**	99
Insulin	Anormal (374) zero values	768	30.5	79.799	115.2	**0**	846
BMI	Anormal (11) zero values	768	32	31.992	7.884	**0**	67.1
DPF	No missing values	768	0.3725	0.4718	0.3313	0.078	2.42
Age	No missing values	768	29	33.240	11.760	21	81
Outcome	Target variable	768	0	0.3489	0.4769	0	1

Although dataset does not explicitly show missing values, missing values are represented as zeros in several features such as BMI and Glucose. As zero BMI it would indicate there is no body mass, which is not realistic. Also, glucose and diastolic blood pressure of 0 is biologically implausible. Same things imply to insulin and skin thickness [20]. These zero values are considered as missing values, several data imputation methods applied such as mean, median, and tree-based model to impute. All preprocessing methods used are described in section preprocessing in methodology [21], [22].

4. METHODOLOGY

This study aims to build a reliable machine learning model that is capable of accurately predicting diabetic cases using eight features (see table (1)). The flowchart (see figure 1) illustrates the procedural pipeline, starting by reading original dataset, then data preprocessing phase aims to enhance data quality, advancing to feature engineering where new features added to the original dataset. Finaly. concluding with cross-validated performance metrics.



.

Figure (1): *Flowchart of the methodology*

4.1 Data Preprocessing

The initial step in the methodology involved conduct several manipulations on the Pima Indian Diabetes (PID) dataset to enhance it and prepare it for machine learning training. This preprocessing phase consisted of three key stages:

4.1.1 Data Exclusion Criteria

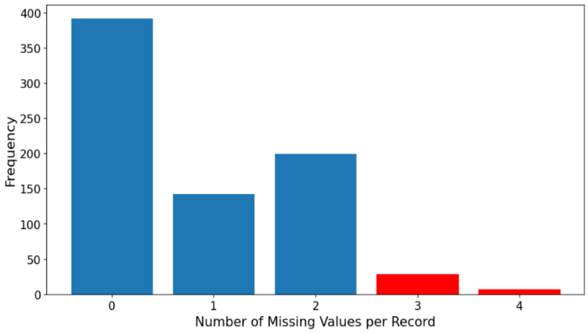


Figure (2): *Frequency of missing values per record*

Given that the dataset contains only 8 features, exclusion criterion adopted by removing patients (rows) with three or more missing values. As the percentage of missingness in three and four features in patient data ranges from 37.5% to 50% which is excessive missing data. These excessively incomplete rows could introduce significant quality concerns. So, to preserve data quality these rows eliminated and only rows with few missing values retained and imputed later. After exclusion criteria applied, samples reduced from 768 to 733. This strict criterion guaranteed that model training will be based on relatively complete data points.

4.1.2 Data Imputation

Data Imputation: To address the issue of missing values in the remaining patients' records (after exclusion criteria applied), an Iterative Imputer was used. This method uses features of the dataset to estimate missing values. Specifically, we implement the Iterative Imputer with a Random Forest Regressor (10 trees) as the estimator, which is particularly effective at capturing non-linear relationships and complex interactions between variables.

4.1.3 Class Imbalance

Class Imbalance: In addition to missing values, the dataset target variable shows a high imbalance issue, most of reported patients are non-diabetic (outcome =0) with percentage 65.1%. This imbalance could negatively affect model training, as many machine learning algorithms (i.e. logistic regress and decision tree) tend to be biased toward the majority class, potentially compromising the performance on the minority non-diabetic class [19]. To address this issue, Synthetic Minority Over-sampling Technique (SMOTE) method used to increase diabetic cases (34.9%) [20]. By adding artificially created data samples for the minority class (diabetic cases). SMOTE could help to mitigates the risk of model bias and improves the predictive performance for both classes.

4.2 Features engineering

To enhance the predictive power of machine learning models for diabetes classification in the Pima Indians Diabetes dataset, a series of feature engineering techniques were developed and implemented. These techniques aimed to capture non-linear relationships, reduce noise, and introduce domain-informed features.

Table (2): *Features importance ranked by Relief method*

	Feature	Chi2	ReliefF	InfoGain
New Feature engineered	Composite_Risk	11.71	0.15	0.21
New Feature engineered	Norm_Glucose	20.77	0.12	0.20
Original Feature	Glucose	1663.26	0.11	0.19
New Feature engineered	BMI_Age	20403.26	0.07	0.16
Original Feature	BMI	84.07	0.06	0.10
Original Feature	Age	286.39	0.06	0.17
Original Feature	DPF	8.67	0.05	0.03
New Feature engineered	Norm_DPF	4.42	0.05	0.03

New Feature engineered	Norm_Age	13.27	0.05	0.19
Original Feature	SkinThickness	166.26	0.04	0.13
New Feature engineered	Norm_BMI	3.78	0.04	0.10
New Feature engineered	Age_BMI_Ratio	1.82	0.04	0.06
New Feature engineered	Pregnancies_Age	10838.81	0.04	0.11
Original Feature	BloodPressure	44.87	0.04	0.12
Original Feature	Pregnancies	171.51	0.03	0.19
New Feature engineered	Insulin_Normal	12.48	0.02	0.09
New Feature engineered	Age_Group_Encoded	50.78	0.01	0.08
New Feature engineered	Insulin_High	9.92	0.01	0.06
New Feature engineered	Insulin_Very High	15.21	0.01	0.02

Engineered features were created by transforming the Pima Indians Diabetes dataset's original eight features through interactions (e.g., BMI × Age), ratios (e.g., Age/BMI), binning (e.g., Age categories), normalization (e.g., min-max scaling), composite scoring (e.g., weighted risk sum), and rule-based thresholds (e.g., Glucose > 100), enhancing model input. To evaluate the relevance of all features in the training dataset Chi-Square, Relief, and Information Gain used.

Table (2) shows features importance using three methods (Chi-Square, Relief, and Information Gain) sorted by Relief. Features importance indicate that new features created have high discrimination power. As many of new engineered features are among top ten features.

4.4 Machine Learning Model Development and Finetuning

Following data preprocessing, several light-weight machine learning algorithms were trained using the preprocessed Pima diabetes dataset. Algorithms used are Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, Artificial Neural Networks (ANN) and Logistic Regression. All Machine learning models implemented and trained using scikit-learn library and python programming language.

5. RESULTS AND DISSCUSSION

Experiments conducted to assess the performance of models for models mentioned in table 2. Dataset used for the training and evaluation is Pima Indians Diabetes dataset after several preprocessing methods applied to improve data quality and impute missing values and it worth to mention that rows with three or more missing values were excluded. As dataset size is relatively small stratified cross-validation (5 folds) used for sampling, to ensure there is enough data per fold for evaluation that reflect real model performance. The results are summarized in Table 1, which reports key performance metrics such as accuracy, F1-score, ROC-AUC, recall, and precision.

Table (3): *Results of 5-Fold Stratified Cross-Validation (no features engineering and exclusion criteria applied)*

Model	**AUC**	**Accuracy**	**F1**	**Precision**	**Recall**	**MCC**
Random Forest	**0.861**	0.785	**0.78**	0.799	0.762	0.571
SVM	**0.861**	0.782	0.778	0.794	0.762	0.564
KNN	0.85	0.777	0.772	0.789	0.756	0.554
Gradient Boosting	0.851	0.774	0.769	0.787	0.752	0.549
ANN	0.821	0.766	0.76	0.779	0.742	0.533
Logistic Regression	0.843	0.761	0.755	0.774	0.738	0.523

Random Forest achieved highest performance (78% F1 score), this show indication that random forest demonstrates tolerance to data noise. Following the initial results obtained using preliminary processing stage (addressing zeros values and oversampling minority class), a series of data enhancement method implemented. To evaluate the proposed data preprocessing methods applied, models were re-applied with same hyper-parameters settings and compared with the prior results.

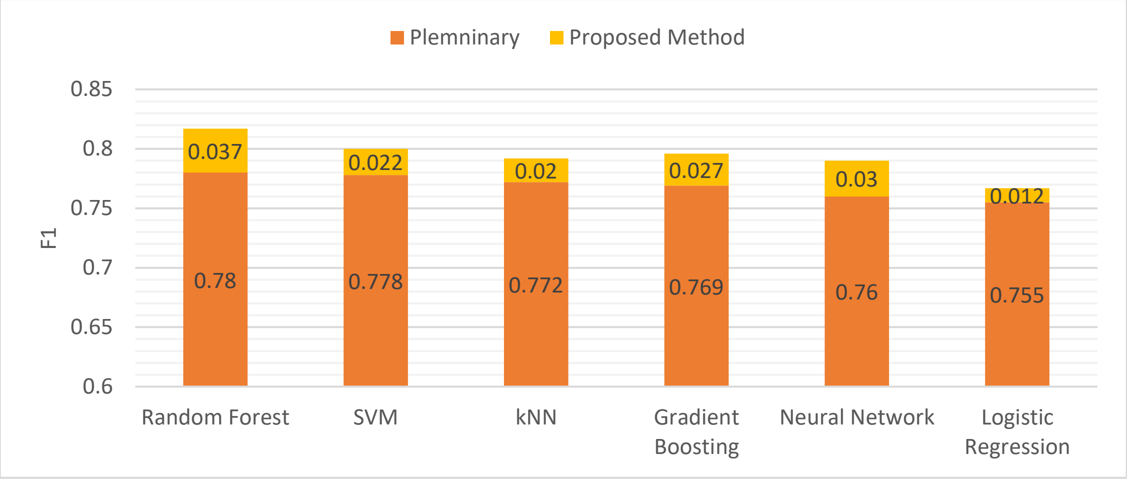


Figure (4): *comparison between several machine learning models results and enhancements*

The proposed improvements, involving the exclusion of data points with high missingness and the new engineered features, show performance enhanced averagely by 2.5% across all models. Random Forest’s out form 78% in terms of F1-score, also got the highest enhancement by 3.7%.

The results demonstrate that ensemble learning (Random Forest) , are highly effective for diabetes prediction on the Pima dataset even before the proposed enhancement and preprocessing operations. Random forest effectiveness in capturing non-linear relationships and mitigating the impact outliers and noise.

6. CONCLUSIONS AND FUTURE WORK

The findings of this study demonstrate the significant impact of data preprocessing and feature engineering on the performance of machine learning models for diabetes prediction using the Pima Indians Diabetes Dataset. By removing rows with excessive missing values, we ensured that the dataset used for training was more reliable and representative of real-world scenarios. This step not only improved the quality of the input data but also enhanced the realism of synthetic samples generated by SMOTE oversampling, as the algorithm operated on a cleaner and more consistent dataset. Furthermore, the introduction of new engineered features—such as interaction terms, ratio features, non-linear transformations, composite risk scores, and rule-based features—provided additional predictive power by capturing complex relationships and domain-specific insights that were not explicitly modeled in the original feature set.

The results highlight the superiority of bagging ensemble learning, particularly the Random Forest, which achieved the highest performance across multiple metrics. Performance enhancement emphasizes the importance of preprocessing (data imputation and oversampling) and feature engineering to address challenges such as class imbalance, missing data, and limited feature diversity. The enhanced model performance demonstrates the potential of machine learning to support early detection and intervention in diabetes management, ultimately contributing to improved healthcare outcomes.

The small size and biasness of Pima dataset restricts model generalization power, for exploring methods such as federated learning to enlarge dataset even without data being shared. Additionally, involve more features such as lifestyle information and wearable device data, could show further enhancement of models’ performanc

REFERENCES

[1] International Diabetes Federation (IDF). (2021). Diabetes Atlas, 10th Edition.

[2] World Health Organization (WHO). (2021). Global Report on Diabetes.

[3] American Diabetes Association (ADA), "Diagnosis and Classification of Diabetes Mellitus," Diabetes Care, vol. 37, no. 1, pp. S81–S90, 2023.

[4] WHO. (2020). Global Estimates of Undiagnosed Diabetes.

[5] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swani, S. M., Blau, H. M., ... & Thierauf, B. S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[6] Obermeyer, Z., Powers, B. J., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

[7] Beam, A. L., & Kohane, I. S. (2016). Big data and machine learning in health care. Jama, 316(21), 2363-2364.

[8] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4), 230-243.

[9] https://archive.ics.uci.edu/datasets/

[10] Abnoosian K., et al., (2023), "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models", https://doi.org/10.1186/s12859-023-05465-z

[11] Viswanatha V., et al., (2023), "Diabetes Prediction Using Machine Learning Approach", DOI: 10.37896/sr10.8/008

[12] SAIHOOD Q., et al, (2023), "A practical framework for early detection of diabetes using A practical framework for early detection of diabetes using ensemble machine learning models"

[13] Ganie S., et al, (2023), "An ensemble learning approach for diabetes prediction using boosting techniques", https://doi.org/10.3389/fgene.2023.1252159.

[14] Reza M., et al., (2024), "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data", https://doi.org/10.1016/j.heliyon.2024.e24536.

[15] Kalyani K., et al, (2024), "Diabetes Prediction Using Random Forest".

[16] Khan Q., et al, (2024), "An intelligent diabetes classification and perception framework based on ensemble and deep learning method PIMA Dataset", DOI 10.7717/peerj-cs.1914.

[17] Maryam I Mousa Al-Khuzaay, Waleed A Mahmoud Al-Jawher "New Proposed Mixed Transforms: CAW and FAW and Their Application in Medical Image Classification" International Journal of Innovative Computing, Volume 13, Issue 1-2, Pages 15-21, 2022.

[18] Salih M.,et al, (2024), "Diabetic Prediction based on Machine Learning Using PIMA Indian Dataset".

[19] Hamid M Hasan, AL Jouhar, Majid A Alwan "Face recognition using improved FFT based radon by PSO and PCA techniques" International Journal of Image Processing (IJIP), Volume 6, Issue 1, Pages 26-37, 2012.

[20] American Diabetes Association (ADA), "Diagnosis and Classification of Diabetes Mellitus," Diabetes Care, vol. 37, no. 1, pp. S81–S90, 2023.

[21] Rasha Ali Dihin, Ebtesam AlShemmary, Waleed Al-Jawher "Diabetic Retinopathy Classification Using Swin Transformer with Multi Wavelet" Journal of Kufa for Mathematics and Computer, Volume 10, Issue 2, Pages 167-172, 2023.

[22] AHM Al-Heladi, WA Mahmoud, HA Hali, AF Fadhel "Multispectral Image Fusion using Walidlet Transform" Advances in Modelling and Analysis B, Volume 52, Issue 1-2, Pages 1-20, 2009.

[23] Japkowicz, N., & Stephen, S. "The class imbalance problem: A systematic study," Intelligent Data Analysis, vol. 6, no. 5, pp. 429–449, 2002.

[24] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[25] Waleed A Mahmoud Al-Jawher, Sarah H Awad "A proposed brain tumor detection algorithm using Multi wavelet Transform (MWT)" Materials Today: Proceedings, Volume 65, Pages 2731-2737, 2022.