Revised study: Enhancing Early Diabetes Prediction Using Machine Learning - An Ethical and Practical Approach

Student ID: 23220052

Abstract— In recent years, the healthcare industry has witnessed a rise in innovative solutions through the development of artificial intelligence (AI). This led to a revolutionary shift in medical practices and chronic conditions diagnosis such as diabetes. Diabetes, a global condition, results in severe health complications if not managed properly, making early prediction and intervention crucial to enhance patient outcomes. Our project, which is an additional study to our previous work, uses supervised learning on a Kaggle dataset containing health and lifestyle indicators. This approach consists in rigorous data analysis between diabetes and crucial variables such as blood glucose levels, HbA1c levels, age, BMI, gender and lifestyle habits to develop a Machine Learning (ML) model that predicts diabetes in patients with high accuracy on an unseen dataset. Finally, we discuss our results, emphasizing on the ethical considerations of inclusiveness and equity in healthcare, aiming to advance preventive healthcare strategies and promote better health outcomes globally.

Keywords: Machine Learning, Healthcare, Artificial intelligence, Diabetes, Predictive analysis

I. Introduction

The healthcare industry has witnessed a rise in AI-driven innovations. This led to a revolutionary shift in medical practices, diagnosis and patient-centric care [1]. Indeed, complex algorithms based on Machine Learning (ML) become increasingly integrated to clinical decision-making processes [2]. By providing accurate, efficient and scalable solutions, these AI technologies are critical to address healthcare challenges such as early disease detection. Traditionally, diagnostic methods rely on symptom manifestation, which often lead to delayed and incorrect diagnosis. However, through ML models, predictive analysis leverages large patient datasets to identify early disease patterns and risk factors [3]. This allows timely intervention, improved patient outcomes and survival rates, which are crucial for conditions like diabetes which have better prognoses and treatment success rates when identified early [4]. In 2023, diabetes affects over half a billion people globally [5] with 1.5 million deaths annually [6].

Diabetes, which happens to be a chronic disease, is identified by the increased levels of sugar in blood due to patient inability to produce or make use of insulin [7]. This metabolic disorder manifest in 2 types: Type 1, an autoimmune disease; and Type 2, related to lifestyle habits such as physical activity, diet, obesity, smoking and alcohol consumption [8]. However, diabetes symptoms usually go without notice since they come with excessive urination, over fatigue as well as blurry vision [9]. This may lead to slower diagnosis and complications such as Cardiovascular diseases (CVDs), retinopathy and Alzheimer's Disease [10]. These complications, arising from the disease lack of management, position diabetes as a significant public health challenge, with a projection to double to 1.3 billion people by 2050 [11]. However, given that its early detection is essential for effective management and preventing complications, it is equally of great importance to use AI in developing predictive models through ML methods.

Supervised learning models, a category of ML, use labelled data to train algorithms that can classify or predict results from new input data. Some examples include Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), and k-Nearest Neighbors (k-NN) [12]. Such models are able to identify high-risk patients, predict disease risks and advise on preventive measures that will lead to patient success and lower healthcare costs by cutting the need for intense treatment and hospitalizations [13].

In this revised study, we will explore different predictive model for diabetes, develop our own, outline the methodology used, describe and discuss our results as well as shed light on the ethical implications of our work.

II. Literature review

Effectively managing diabetes is a significant public health challenge [14], that has led to various research focusing on its early detection and prediction through ML algorithms. These approaches analyze complex data patterns to predict high-risk individuals based on risk factors and lifestyle habits. LR is a statistical and binary classification method used to estimate the probability of an outcome based on input features [15]. In terms of diabetes prediction, LR models have been widely used due to their simplicity and interpretability [16]. Indeed, Nguyen et al. (2019) used LR to predict diabetes risk using features such as age, BMI, and blood glucose levels [17]. The model achieved high accuracy and performance, and was particularly acknowledged for its simple interpretation. It was found that each feature was a significant predictor of diabetes, allowing better understanding of their impact on diabetes risk. Moreover, according to Rajendra and Latifi (2021), LR was also considered as an efficient algorithm in building prediction models [18]. However, its accuracy did not only depend on the chosen algorithm but also on data pre-processing. Indeed, removing duplications and null values, and using cross-validation was essential to improve its performance. Likewise, Anderson et al. (2015) raised concerns about the poor development of predictive models due to inappropriate selection of covariates, missing data, and small sample sizes [20] . Furthermore, the reliability and quality of these predictive

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

models highlighted significant variation based on geography, available data, and ethnicity [17]. Indeed, LR performance presents limitations when dealing with complex, non-linear relationships in the data as it may not capture intricate patterns as effectively as more sophisticated algorithms [21]. Despite these limitations, LR was found to classify high-risk individuals accurately [19] and assist in diabetes prevention and management when validated against DT. Indeed, Joshi and Dhakal (2021) study compared the LR method with the DT method and found that the risk factors for diabetes identified by LR were validated by DT, suggesting that this model can help classify high-risk individuals accurately and help in the prevention, diagnosis, and management of diabetes [22]. DTs are intuitive models that create a tree-like structure by separating data into branches based on feature values [23]. Being fast, easy to interpret and performing well even on large dataset, they learn simple decision rules provided by data features, handle well non-linear relationships and develop models that predicts the target value [24]. Indeed, Iparraguirre-Villanueva et al. (2023) study demonstrated the application of DTs in diabetes prediction, showing the model's ability to manage complex interactions between BMI, insulin levels, and age [25]. However, according to Smolic (2023) these models are also prone to overfitting, especially when they are deep and complex [26]. RF, which develop multiple DTs and combine their predictions [27], offer a solution to this problem by reducing overfitting and handling complex interactions between features more effectively. Indeed, in terms of predictive accuracy, this approach outperforms LR due to its ability to handle complex interactions between features, as well as DTs due to its ensemble approach that reduces overfitting [28]. Studies by Wang (2024) and Mahboob Alam, et al. (2019) used RF on diabetes datasets, resulting in high accuracy[29][30], while Ahmed et al. (2021) study compared RF approach to others, and highlighted it having the highest accuracy [31] due to its effectiveness in identifying diabetes predictors, managing data imbalances and handling large datasets [29], as well as its ability to provide a balance between accuracy and interpretability [30]. However, despite its strengths, RF models can be computationally intensive and less interpretable than simpler models like LR [32]. Moreover, SVMs have been found to perform well in diabetes prediction, offering high accuracy and robustness. Indeed, Sharma and Shah (2021) study compared LR, DT, RF and SVMs methods for diabetes prediction and discovered that SVMs performed best [33], offering the highest accuracy and robustness [34]. However, according to GeeksforGeeks (2023), SVM method require complete and small dataset as they cannot handle missing values, consume a lot of memory and are very slow if many feature are within the dataset [35]. Furthermore, recent studies have also explored the use of DL models such as CNNs, ANNs, and k-NN in diabetes prediction. ANNs have also been effective in diabetes prediction. Alehegn et al., (2017) study reported accuracies up to 88.41% [36]. Indeed, their ability to handle non-linear data and their adaptability to various input features make them a suitable choice for developing decision support systems for diabetes prediction [36]. According to Zhou et al., (2018), regardless of requiring substantial computational power and data preprocessing, the ANN model achieved the highest accuracy (85.46%) in predicting diabetes mellitus, highlighting the importance of considering multiple factors in disease prediction [37]. Despite their similar limitation, CNNs, known for their deep architecture and high-level feature representation, have shown promise in diabetes prediction [38]. Indeed, according to Swapnaet al.,(2018), they have achieved high accuracy (93.6%) in predicting blood glucose levels and classifying patients as euglycemic, hypoglycemic, or hyperglycemic [39]. Due to their ability to convert numerical data into images, these models have demonstrated robustness in early diabetes diagnosis, achieving high prediction accuracies [38]. Furthermore, k-NN algorithms, which classifies data points based on their proximity to other, has also been applied to diabetes prediction [40]. While not as commonly reported as CNNs and ANNs, this non-parametric powerful ML algorithm offers simplicity and interpretability, making it a viable option to predict and classify diseases based on patient data [41]. Indeed, in Muthu et al., (2023) study, k-NN was used to predict the risk of type 2 diabetes using a dataset of electronic medical records from patients diagnosed with diabetes and healthy individuals [40]. The model's performance was assessed using accuracy, precision, recall, and F1-score, demonstrating its high accuracy (70%) and effectiveness in classifying patients at high and low risk of developing diabetes [40]. Finally, ensemble techniques have shown substantial promise in enhancing the accuracy of diabetes prediction models [42]. Indeed, Edeh et al. (2022) study found that while SVM achieved the highest accuracy (83.1%) among individual models, ensemble methods consistently provided better performance [43]. Similarly, Chen et al. (2017) enhanced prediction accuracy to 90.04% by preprocessing data with the k-means algorithm for data reduction before applying a decision tree classifier [44]. This highlights the efficacy of combining data preprocessing techniques with ensemble learning. Moreover, according Alehegn et al., (2017) study, an ensemble approach was implemented to enhance prediction accuracy, achieving an accuracy of 78.74%, surpassing individual models [36]. However, despite these successes, some studies (as previously mentioned) suggest that individual models, can still offer competitive performance. This might be due to the complexity and computational intensity of ensemble models which pose challenges and necessitate further research to refine these techniques for practical, real-world applications [45]. Therefore, the need for high-quality data, efficient feature selection, and robust preprocessing methods is crucial for optimizing the performance of ensemble models in predicting diabetes [46].

III. METHODOLOGY

A. Our dataset

The "diabetes_prediction_dataset.csv file" used for our ML model development was retrieved from Kaggle [35]. It comprises 100,000 patients medical and demographic data, including essential features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood glucose level and their diabetes status. These factors are vital to predicts diabetes in patients, as it allows Healthcare professionals (HCPs) to identify patients at risk, and understand the correlation between these factors and diabetes risk. Therefore, our project aims to train our ML model on this labelled dataset with known outcome, to accurately predict patients at risk of diabetes.

Features	Description
Gender	There are three categories in it male, female and other (59% Female, 41% Male). The biological sex of the individual can have an impact on their susceptibility to diabetes.
Age	Age ranges from 0-80 in our dataset. It is an important factor as diabetes is more commonly diagnosed in older adults.
Hypertension	Hypertension values are 0 if the patient has no hypertension condition, and 1 if they do. When a patient has diabetes, his blood pressure in the arteries is persistently elevated.
Heart disease	Cardiovascular diseases values are 0 and 1, where 0 indicates no heart diseases and 1 indicates that the patients have heart diseases. Patients with CVDs are associated with an increased risk of developing diabetes.
Smoking history	Smoking history consists of 5 categories: not current, former, No Info, current, never and ever. It is considered a risk factor for diabetes and can exacerbate the complications associated with the disease.
Body Mass Index (BMI)	The range of BMI in our dataset is from 10.16 to 71.55. Please note that BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese. BMI consists of the measure of body fat based on weight and height. A high BMI value is linked to a higher risk of diabetes.
Haemoglobin A1c (HbA1c)	HbA1c level measure the patient's average blood sugar level over the past 2-3 months. Indeed, Glycated haemoglobin is a form of haemoglobin that is chemically linked to a sugar. High levels of HbA1c (more than 6.5%) usually indicate a greater risk of developing diabetes.
Blood glucose level	Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes.
Diabetes	Diabetes feature is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

Figure 1: *Dataset features description*

B. Data Exploration and Pre-processing

Before starting our project, we first import all required libraries for our program solution.

In order to predict diabetes, we first import and load our dataset from the 'diabetes_prediction_dataset.csv' file. Then, we thought it would be beneficial to generate a summary of our dataset to gain insights into the data structure, distribution, inconsistencies and missing values. Such insights allow us to understand the overall behavior and spread of our dataset, guiding our selection of appropriate analysis techniques and models. Once done, our dataset undergoes pre-processing to ensure uniformity and compatibility for ML models. Given the presence of both numerical and non-numerical data, we convert all non-numerical data into numerical equivalents, to ensure consistent handling of our data. Then, once our dataset is pre-processed and cleaned, we split it into training (70%), testing (25%) and deployment (5%) sets which we will use to train, evaluate and validate our model. However, prior to developing our model, we identified and visualized how many people have diabetes in our training and testing datasets, as well as checked for any data imbalances. This step is crucial as it ensures that ML models perform well across all classes, avoiding bias towards the majority class and thereby providing more accurate and reliable predictions. Finally, we performed a correlation analysis to understand the relationship between numerical features and the 'diabetes' target variable in our dataset. This way, we will be able to identify which features have the most significant impact on diabetes prediction.

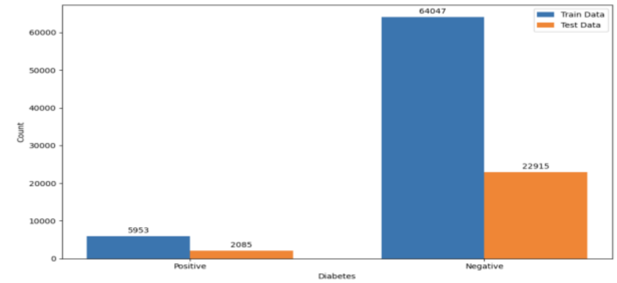

Figure 2: *Counts of Positive and Negative Entries in Train and Test Data*

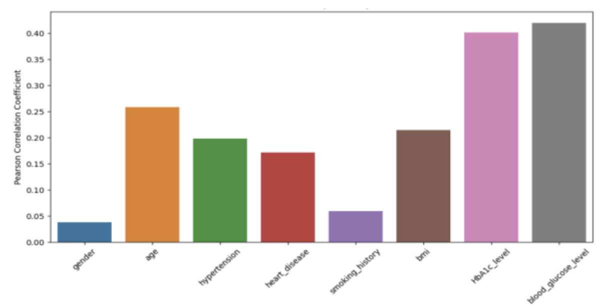

Figure 3: *Correlation Coefficients (Pearson) with Diabetes*

C. Classifiers Training, Evaluation, and Comparison

From our previous study and based on our literature review, we decided to evaluate the effectiveness of LR, DT, and RF models in predicting diabetes. We compared their accuracy to determine the most suitable model for our training. By doing so, RF classifier showed the highest accuracy in predicting diabetes based on our dataset. Therefore, our RF model was trained on the training data after which its performance was evaluated using testing data. Using the GridSearchCV function, we identified the best combination of parameters for our model and fine-tuned it to enhance its accuracy on the testing data. We then re-evaluated the model using the optimized hyperparameters to further improve its performance. Please note that we generated classification reports, providing precision, recall, and F1-score metrics as well as a confusion matrix before and after hyper-tuning our RF model. This was done to understand how well the model predicts each category and assess the changes in model performance in terms of true positives, true negatives, false positives, and false negatives. This way, we were able to understand the improvements made by hyperparameter tuning and ensured that our model provides reliable predictions for all classes. Subsequently, we performed a cross-validation on our RF model across 5 folds of the training data. This step helped mitigate the risk of overfitting and ensures the model generalizes well to new data. After finalizing the model, we saved it for future use, allowing for predictions on new data without retraining from scratch. We then deployed our model on unseen deployment data (5%, X_deploy, y_deploy) to assess its performance on completely new and independent data, validating its robustness and reliability before making critical decisions. Finally, we calculated the accuracy score of our model on the deployment dataset to validate and maintain its efficacy in real-world applications. We also analyzed the dataset features to identify those with the most significant impact on our model's predictions.

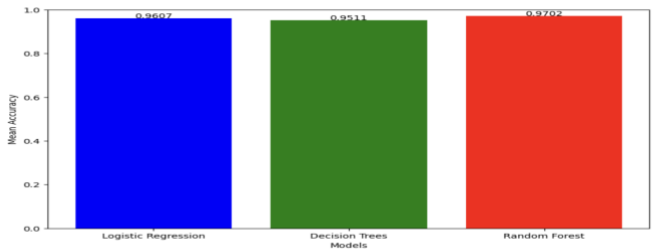

Figure 4: *Accuracies of RF, DT and LR models*

```
Confusion Matrix before hypertuning our RF model:
 [[22828    87]
 [  660  1425]]
```

```
Confusion Matrix after hypertuning our RF model:
 [[22913     2]
 [  695  1390]]
```

Figure 5: *Matrix of RF before and after hyper-tuning*

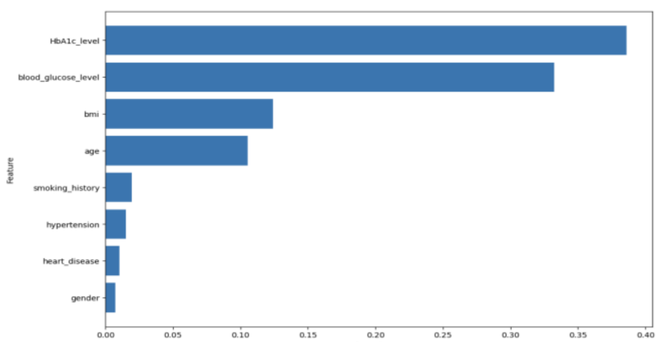

Figure 6: *Feature importance*

For this study, we selected SVM, k-NN, CNN, and ANN supervised learning ML models, and compared their accuracies to determine the most suitable one for our model training. We began by training and evaluating SVM and k-NN models using cross-validation. This involved setting up a preprocessing pipeline with SimpleImputer to handle missing values and StandardScaler to standardize features. We then initialized a dictionary of classifiers, including SVM and k-NN, and created pipelines combining the preprocessor and classifiers. Each pipeline was evaluated using 5-fold cross-validation, computing and printing accuracy scores. Then, we evaluated our CNN model, by standardizing the feature data and reshaping it to meet the model input requirements. We defined our CNN model with two 1D convolutional layers, a flattening layer, dense layers including dropout for regularization, and compiled the model using the Adam optimizer and categorical cross-entropy loss function. The model was trained over 10 epochs with a batch size of 32, its performance was evaluated on the test data and we printed its accuracy. Finally, we prepared our ANN model by standardizing the features. Our ANN model was defined with two dense layers and dropout for regularization, compiled with the Adam optimizer and categorical cross-entropy loss function, and trained over 10 epochs. The model's performance was evaluated on the test set and we printed its accuracy.

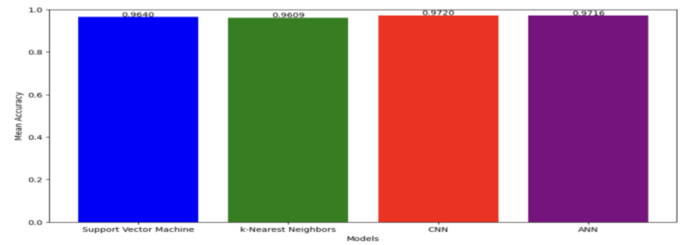

Figure 7: *Accuracies of SVM, kNN, CNN and ANN models*

D. Best Model: CNN

From our evaluation, our CNN model performed best, scoring an average accuracy of 97.20%. As a result, we chose our CNN model for final training, testing, and deployment to assess how well it performs on unseen data. To do that, we first trained our CNN model on our preprocessed training data. The second step consisted of standardizing the features and reshaping them into the dimensions needed for our model. Then, we fit them into our model with two 1D convolutional layers, a flattening layer, dense layers with dropout to avoid overfitting, and finally a softmax output layer. It was then compiled using the Adam optimizer and categorical cross-entropy loss function, trained for 10 epochs with a batch size of 32. Afterward, we evaluated our model performance on our testing data. For this, we generated predictions, a classification report including precision, recall, F1-score, and support for each class, as well as a confusion matrix showing how many true positive, true negative, false positive, and false negative predictions were predicted. This provided and evaluation of our model performance before hyper-tuning as well as insights into our model's accuracy and error distribution across different classes. Following this, hyperparameter tuning was performed to find the best configuration for our CNN model. The process involved defining a grid of hyperparameters, including batch size, epochs, optimizer, and dropout rate, and iterating it through all possible combinations to find the model with the highest validation accuracy. Once done, we conducted a 5-fold cross-validation to ensure our models robust performance. This involved training and validating the model across different subsets of the data to mitigate the risk of overfitting and ensure generalizability. We again generated predictions, a classification report that includes precision, recall, F1-score, and support for each class, as well as a confusion matrix. This provided a comprehensive evaluation of our model's performance after hyperparameter tuning and offered insights into its accuracy and error distribution across different classes, allowing us to compare its performance and confirm the improvements from hyperparameter tuning.

```
Confusion Matrix before hypertuning our CNN model:
 [[22907     8]
 [  691  1394]]
```

```
Confusion Matrix after hypertuning our CNN model:
 [[22911     4]
 [  691  1394]]
```

Figure 8: *CNN confusion matrix before and after hyper tuning*

Finally, our best CNN model was saved as 'best_model2.pkl' for future use, and deployed on our new unseen deployment data (5%). Then, using our optimal CNN model, we generated predictions and saved them to 'deployment_results_with_predictions_cnn.csv'. We calculated our model accuracy to assess how well it performs on completely new and independent data that it hasn't seen during training or evaluation. This is crucial as it provides insights into our model's ability to generalize to real-world scenarios, validating its robustness and reliability before it is used for critical decision-making. Additionally, we also generated a classification report and confusion matrix to evaluate our model performance on deployment data, offering detailed metrics on its predictive accuracy and error distribution.

IV. Results

Initially, our dataset was complete and had no missing values, allowing straightforward use. Upon splitting our dataset, our training data consisting of 70,000 cases, had 5,953 positive and 64,047 negative diabetes cases, with a difference of 58,094. Whereas, our testing dataset of 25,000 cases, had 2,085 positive, and 22,915 negative, resulting in a difference of 20,830. Therefore, our dataset distribution was slightly imbalanced (8.50% positive in training, 8.34% in testing; 91.50% negative in training, 91.66% in testing), which did not affect our model's performance as it achieved high accuracies. Among evaluating all potential models (LR, DT and RF), we noticed that our RF model emerged as the most effective. Indeed, while LR achieved a mean accuracy of 96.07%, and DT 95.11% , RF achieved the highest one (97.02%). Therefore, in our previous study, we selected the RF model for further refinement and deployment due to its superior performance. Once our model trained, our model returned a high training accuracy of 99.92% but maintained almost the same testing accuracy of 97.01%. Indeed, as shown in the classification report, our model yielded a very high precision for class 0 at 0.97 and a recall of 0.68 for class 1 (diabetic), resulting in an overall balanced F1-score of 0.97. On the other hand, according to our models confusion matrix before hyperparameter tuning revealed that it was correct about 22,828 nondiabetic cases and 1,425 diabetic cases; however, it also mislabeled 87 nondiabetic patients as diabetic and 660 diabetic patients as nondiabetic. While the low number of false positives, demonstrated that it is good at not mislabeling nondiabetic cases as diabetic, the moderately high number of missed diabetic cases, showed there is much improvement to do in the our model sensitivity in detecting diabetes. We then fit our RF model to its best parameters, which consisted of a max depth of 10 with a min sample split of 2 and n_estimators of 100. When further evaluating the best-tuned RF mode, it achieved a slightly improved testing accuracy of 97.21%. Subsequently, while cross- validating our results, our model performance resulted in a mean accuracy of 97.02%. Moreover, following hypertuning, our RF model improved its classification metrics substantially. Indeed, it obtained a precision of 0.99 for class 0 with a recall of 1.00 and 0.80 precision for class 1 with a similar recall of 1.00. The confusion matrix further illustrated these improvements with 22,913 true negatives, only 2 false positives, 695 false negatives, and 1,390 true positives. This demonstrates a substantial reduction in misclassifications and highlights the robustness of our hyper-tuned model in accurately predicting diabetes cases. Finally, when using our model on our deployment data, we noticed that our RF still maintained a strong accuracy of 96.88%. When we deployed our hyper-tuned Random Forest model on new, unseen data, the performance remained strong, reflecting its ability to generalize effectively to real-world scenarios. The model achieved an accuracy of 97% on the deployment dataset. It maintained a precision of 0.97 for class 0 with a perfect recall of 1.00 and a precision of 0.80 with a recall of 0.66 for class 1. Moreover, the confusion matrix for the deployment data showed 4,538 true negatives, 0 false positives, 156 false negatives, and 306 true positives. These results confirmed our model's reliability and its high predictive performance when applied to completely new data.

```
Confusion Matrix on deployment data:
 [[4538    0]
 [ 156  306]]
```

Figure 9: *Confusion matrix of RF model on deployment data*

Moreover, in our RF model, the feature importance rankings were determined to understand which variables most significantly influence the prediction of diabetes. The feature with the highest importance is HbA1c_level, indicating that glycated hemoglobin levels are the most critical factor in predicting diabetes. Following HbA1c_level, blood_glucose_level stands as the second most important feature, underscoring the role of blood sugar levels in diabetes diagnosis. BMI (Body Mass Index) is the third most influential feature, highlighting the impact of body weight and obesity. Age is the fourth important feature, reflecting the age-related risk of diabetes. Smoking_history comes next, suggesting that smoking status also contributes to diabetes risk. Hypertension follows, indicating a link between high blood pressure and diabetes. Heart_disease is the penultimate feature, emphasizing the association between cardiovascular conditions and diabetes. Finally, gender is the least important feature in our model, suggesting that while there may be differences in diabetes risk between males and females, they are less critical compared to the other features.

In this study, we compared the SVM model which exhibited strong performance, resulting in a mean accuracy of 0.9640. This indicates a consistent and robust performance across different subsets of the dataset. Similarly, the k-NN model demonstrated commendable results, with a mean accuracy of 0.9609. While both models performed well, the ANN model showed slightly superior accuracy of 97.60%. However, despite these strong performances, the CNN model emerged as the top performer, with a test accuracy of 97.20%. This was further affirmed by the average accuracy of 97.19% across different tests, underpinning its reliability. In fact, the first training of our CNN model achieved an accuracy of 97.20% with a precision of 0.97 and a recall of 1.00 for class 0 and 0.99 with a recall of 0.67 for class 1. Indeed, it had 22,907 correctly identified non-diabetic cases and 1,394 diabetic cases. However, there were 8 instances of non-diabetic cases misclassified as diabetic, and 691 diabetic cases classified as non-diabetic. Although there is a slight increase in the false negatives as compared to the pre-hyper-tuning model, the large reduction in false positives shows that our hyper-tuned model is much more precise and efficient in differentiating between diabetic and non-diabetic cases with an overall high accuracy. After hyper-tuning our CNN model with optimal hyperparameters (batch size: 32, epochs: 20, optimizer: 'adam', dropout rate: 0.3), the performance improved slightly. Indeed, the training and testing accuracy further improved up to 97. 22%. The new classification report revealed a precision equal to 0. 97 with a recall of 1.00 for class 0 and 0.98 with a recall of 0.68 for class 1. On the other hand, the matrix showed that the model correctly classified 22,911 individuals as non-diabetic and 1,394 as diabetic. However, for non-diabetic cases 4 of them have been predicted as diabetic and 691 diabetic cases have been predicted as non-diabetic. This modified confusion matrix reveals a relatively slight reduction in false positive and false negative cases when compared to our prior assessment, indicating enhancement in the model's classification accuracy and its ability to reliably differentiate between diabetic and non-diabetic patients. When our hyper-tuned CNN model was deployed on new data, it achieved an accuracy of 96.96%, maintaining its robustness in real-world scenarios. The classification report for the deployment data showed a precision of 0.97 for class 0 with a recall of 1.00 and 1.00 with a recall of 0.67 for class 1. This indicates that our model correctly classified 4,537 non-diabetic cases and 311 diabetic cases. However, there was only 1 instance where a non-diabetic case was incorrectly classified as diabetic, and 151 diabetic cases were misclassified as non-diabetic. Therefore, this matrix highlights the model's strong ability in differentiating between classes on deployment data, with very few misclassifications, hence proving to be robust and reliable in real-world scenarios.

```
Confusion Matrix on deployment data:
 [[4537    1]
 [ 151  311]]
```

Figure 10: *Confusion matrix of CNN model on deployment data*

V. Discussion

A. Discussion of results:

The use of ML methods in our diabetes prediction study resulted into promising outcomes. Both our training and testing datasets exhibited a similar distribution of diabetes cases, indicating consistent representation of the target variable. Before hyperparameter tuning, our RF model showed an extremely high of 99.92% indicating that it fit the training data very well, nearly to the point of perfection, and a slightly lower testing accuracy of 97.01%. On the other hand, our CNN model exhibited a high training accuracy, slightly lower than the RF, and a testing accuracy of 97.20%. Comparing the pre-tuning confusion matrices of our RF and CNN models reveals differences in their performance in classifying diabetic and non-diabetic cases. Indeed, our RF model precisely classified 22,828 non-diabetic cases as truly negative and 1,425 diabetic cases as truly positive. In contrast, it misclassified 87 non-diabetic cases as diabetic and 660 diabetic cases as non-diabetic. However, for our CNN model, there were 22,907 non-diabetic cases and 1,394 diabetic cases predicted correctly, with only 8 false positives and 691 false negatives. For the classification metrics, our RF model precision was 0.97 and recall was 0.68 for class 0, while our CNN model turned in a better precision of 0.99 and a recall of 0.67 for class 1. This comparison indicates that, prior to tuning, our CNN model demonstrated a higher true negative rate and, correspondingly, a lower false positive rate than our RF model, which shows greater precision in non-diabetic cases identification. However, the RF model had a higher true positive rate and lower false negative rate, showing greater sensitivity in detecting diabetic cases. Thus, while the CNN model showed better precision for non-diabetic predictions, the RF model excelled in identifying diabetic instances more accurately. In terms of confusion matrices post hyperparameter tuning, our RF model's testing accuracy improved slightly to 97.21% with a confusion matrix showing 22,913 true negatives, 1,390 true positives, 2 false positives, and 695 false negatives. In fact, it showed a very low number of false positives, which is beneficial in minimizing the number of non-diabetic cases incorrectly classified as diabetic. On the other hand, our CNN model training accuracy slightly increased to 97.22%, meaning that it indicates effective optimization of the model performance, with its confusion matrix showing lightly higher number of false positives (4) but fewer false negatives (691) than the RF model, suggesting better sensitivity in detecting diabetic cases. It can also be noted that the true positive for our CNN model was slightly higher at 1394; implying that it correctly identified more diabetic patients. Concerning the classification report, in our RF model, the precision was 0.97 for class 0 and 1.00 for class 1, while recall was 1.00 for class 0 and 0.67 for class 1, hence resulting in an F1-score of 0.99 for class 0 and 0.80 for class 1. Our CNN model, similarly, exhibited a slightly improved precision for class 1 (0.98) and maintained high recall for class 0. Indeed, it demonstrated similar precision for non-diabetic cases and slightly lower precision for diabetic cases, but with a slight improvement in recall for diabetic cases compared to the RF model. Upon comparing the accuracies of the RF and CNN models on deployment data, we noticed that they both consisted of a robust performance in real-world scenarios. IIn fact, our RF model achieved a slightly lower accuracy of 96.88%, demonstrating its strong capability to generalize effectively to new, unseen data. Similarly, the CNN model maintained a high accuracy of 96.96% on the deployment dataset, reflecting its reliability and robustness in practical applications. Indeed, they both exhibited comparable accuracy levels on deployment data, which underscores their effectiveness in predicting diabetic and non-diabetic cases, with the CNN model slightly outperforming the RF model by a narrow margin. Regarding the confusion matrices of the RF and CNN models on deployment data, both models exhibited strong performance. The RF model correctly identified 4,538 non-diabetic cases and had 0 false positives, demonstrating exceptional reliability in avoiding misclassification of non-diabetic cases. However, it missed 156 diabetic cases, identifying 306 true positives. Conversely, the CNN model correctly classified 4,537 non-diabetic cases with a single false positive, slightly less reliable than the RF model in this aspect. Notably, the CNN model performed better than our RF in terms of sensitivity, returning only 151 false negatives against the 311 true positives, indicating its higher ability to detect diabetic cases. In terms of classification reports, our RF model's showed precision values for class 0 of 0.97 and for class 1 of 1.00, with their recall values remaining the same at 1.00 and 0.66, respectively. Our CNN model maintained the same performance, highlighting how robust it is. Therefore, although our RF model is more suitable when minimizing false positives is very critical in an application, our CNN model has advantages in scenarios where sensitivity to diabetic cases is required. This reflects the model's balanced and robust performance in real-world deployment. However, the choice between these models depends on the requirements of the application— such as the importance of minimizing false positives versus maximizing the detection of true positives.

B. Ethical considerations:

Both our RF and CNN models have demonstrated robust performance in identifying diabetes cases with high precision and recall on new, unseen data, underscoring their potential for practical deployment in healthcare settings. Despite this, each model has its own set of ethical implications, limitations, and areas for improvement. Our RF model has proven effective with high precision and recall, yet it faces challenges related to encoding categorical variables, its inability to remove duplicates, and the absence of ethnicity data. These limitations could introduce inaccuracies and biases, reduce the model's inclusiveness and affect our prediction reliability [48] [49], as well as limit our understanding of diabetes prevalence across different demographic groups [50]. Moreover, factors like genetics and alcohol consumption, not included in our dataset, significantly influence diabetes risk, which potentially introduces inaccuracies in our predictions [8]. Therefore, it is crucial to have more rigorous data pre-processing techniques to ensure quality and integrity, incorporate categorical data more effectively, and include broader demographic data [50]. Addressing these issues is crucial to ensure that our model provides equitable healthcare predictions across diverse populations and avoid reinforcing existing disparities. On the other hand, our CNN model, which achieved impressive accuracy rates, utilizes a novel approach of converting numerical data into images, enabling the use of advanced CNN architectures [51]. This innovative method allows the CNN to process data more effectively but also introduces potential biases during the data conversion process which may lead to the loss or misrepresentation of critical features [51]. Indeed, despite its strong performance in distinguishing between diabetic and non-diabetic cases, our CNN model faced challenges, particularly in its ability to minimize false negatives and its consistent misclassification of some diabetic cases as non-diabetic, leading to a potential impact on the model's performance, despite efforts to mitigate this through tuning and hyperparameter adjustments. This highlights the need for improved data augmentation and feature selection as it could be critical in a healthcare context where missing diabetic cases can have significant consequences, such as health complication or high hospital readmission rates [52][53]. However, the reliance on selected features and data augmentation highlights a trade-off between complexity and interpretability. Additionally, the absence of ethnicity data in our CNN model reveals limits its ability to generalize across diverse populations, potentially overlooking important demographic variations in diabetes prevalence [54]. Therefore, this gap emphasizes the importance of incorporating a broader range of data to ensure equitable healthcare outcomes and avoid reinforcing existing disparities [54]. To address these limitations and enhance our models performance, several strategies could be employed. For instance, to improve our CNN model's performance, data augmentation techniques could be employed to address the class imbalance, potentially increasing the model's ability to generalize to underrepresented classes [55]. Additionally, exploring a broader range of CNN advanced architectures, such as deeper networks or those incorporating attention mechanisms, could enhance the model's ability to detect subtle patterns in the data [56]. Regularization techniques and hyperparameter optimization, beyond the initial tuning, could further refine the model's sensitivity and specificity [57]. Moreover, incorporating external data sources, transformer-based networks, domain-specific features and expanding the feature set might provide additional context, leading to more accurate and robust predictions [58]. Finally, using ensemble techniques such as combining RF and CNN models, might also enhance accuracy and effectiveness [36].

Our solutions' impact on inclusiveness and diversity requires careful consideration, as predictive models can improve healthcare outcomes by enabling early intervention but may exacerbate disparities if access to technology is unequal [59]. Therefore, proactive efforts are needed to ensure equitable access and considerate deployment of predictive models in healthcare settings. Furthermore, integrating diverse data sources and ensuring transparency in model operation are also essential to maintaining ethical standards. Therefore, proactive efforts to refine these models will ultimately lead to better predictions and more inclusive healthcare solutions [60]. Moreover, safeguarding data privacy and confidentiality is crucial, especially for sensitive health information [61]. Therefore, ensuring compliance with data protection regulations and robust security measures are imperative to mitigate risks of data breaches and unauthorized access [62]. Lastly, relying solely on automated prediction models without human oversight can lead to erroneous conclusions, particularly in complex scenarios, potentially causing HCPs to misinterpret the results and misjudge diagnoses [63]. Therefore, providing clear explanations of our model's operation, limitations, and potential biases is crucial for maintaining transparency. Ultimately, prioritizing ongoing evaluation, refinement, proactive efforts to incorporate diverse data and ensure ethical standards, addressing potential biases, maintaining robust data privacy measures, incorporating diverse data, enhancing model inclusiveness, as well as consistently ensuring transparency in our models operation will contribute to more accurate, equitable, and effective healthcare solutions [64].

VI. CONCLUSION

To conclude, diabetes represents a significant public health challenge, as it impacts over half a billion individuals globally. Our revised study has sought to enhance early diabetes prediction and intervention by using advanced ML models that extend beyond our previous research. By evaluating and comparing the efficacy of our RF and CNN models, our aim was to address existing gaps in predictive methodologies and assess their practical implications in real-world healthcare settings. Our findings validate the high efficacy of both RF and CNN models in predicting diabetes across diverse datasets. While our RF model excelled in sensitivity, achieving a higher true positive rate with fewer false positives, our CNN model exhibited superior precision for non-diabetic cases and slightly improved sensitivity post hyperparameter tuning. These results underscore the potential of both models for early diabetes detection, highlighting their effectiveness as tools for enhancing healthcare outcomes. Their successful application in real-world scenarios reinforces their practical utility in clinical settings, enabling HCPs to identify high-risk individuals and implement preventive measures more effectively. Ensuring that predictive models are developed and deployed with these principles in mind is essential for advancing healthcare practices and promoting better health outcomes globally. Indeed, the application of ML into diabetes prediction is quite a remarkable technical-medical alternative in preventive healthcare, with the potential to reduce disease burden and improve patient quality. However, despite their promising outcomes, continuous monitoring and ongoing refinement are essential for models sustainability and accuracy. Indeed, improvements in data pre-processing, reducing biases, including ethnicity data, insuring that data privacy laws are protected, and making the model understandable are crucial for enhancing model performance and ensuring that fairness in predictions remains a priority. Therefore, future researches need to be devoted to the integration of ensemble learning techniques that will enhance model robustness and generalization, the further characterization of features and data sources, and addressing ethical considerations to maximize the impact of AI-driven healthcare solutions. By adhering to these principles, we can advance healthcare practices, reduce disease burden, and improve patient outcomes globally. In other words, the integration of ML into diabetes prediction represents a significant step forward in preventive healthcare, with the potential to transform disease management and enhance quality of life.

REFERENCES

[1] Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In Artificial Intelligence in Healthcare. https://doi.org/10.1016/B978-0-12-818438-7.00002-2

[2] Iqbal, J., Cortés Jaimes, D. C., Makineni, P., Subramani, S., Hemaida, S., Thugu, T. R., Butt, A. N., Sikto, J. T., Kaur, P., Lak, M. A., Augustine, M., Shahzad, R., & Arain, M. (2023).Reimagining Healthcare: Unleashing the Power of Artificial Intelligence in Medicine. Cureus. https://doi.org/10.7759/cureus.44658

[3] Batko, K. and Ślęzak, A. (2022) The use of Big Data Analytics in Healthcare, Journal of big data. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8733917/ (Accessed: 04 July 2024).

[4] Herman, W.H. et al. (2015) Early detection and treatment of type 2 diabetes reduce cardiovascular morbidity and mortality: A simulation of the results of the Anglo-danish-dutch study of intensive treatment in people with screen-detected diabetes in primary care (addition-Europe), Early Detection and Treatment of Type 2 Diabetes Reduce Cardiovascular Morbidity and Mortality: A Simulation of the Results of the Anglo-Danish-Dutch Study of Intensive Treatment in People With Screen-Detected Diabetes in Primary Care (ADDITION-Europe). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512138/ (Accessed: 04 July 2024).

[5] Institute for Health Metrics and Evaluation (no date) Global diabetes cases to soar from 529 million to 1.3 billion by 2050. Available at: https://www.healthdata.org/news-events/newsroom/news-releases/global-diabetes-cases-soar-529-million-13-billion-2050#:~:text=June%2022%2C%202023%20%E2%80%93%20More%20than,published%20today%20in%20The%20Lancet%20. (Accessed: 04 July 2024).

[6] Diabetes (no date) World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=In%202019%2C%20diabetes%20was%20the,of%20cardiovascular%20deaths%20(1). (Accessed: 04 July 2024).

[7] What is diabetes? (no date) National Institute of Diabetes and Digestive and Kidney Diseases. Available at: https://www.niddk.nih.gov/health-information/diabetes/overview/what-isdiabetes#:~:text=If%20you%20have%20diabetes%2C%20your,to%20some%20types%20of%20cancer. (Accessed: 04 July 2024).

[8] Sami, W. et al. (2017) Effect of diet on type 2 diabetes mellitus: A Review, International journal of health sciences. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426415/(Accessed: 04 July 2024).

[9] Diabetes UK (no date) What are the signs and symptoms of diabetes?, Diabetes UK. Available at: https://www.diabetes.org.uk/diabetes-the-basics/diabetes-symptoms (Accessed: 04 July 2024).

[10] W, E. (no date) Complications of diabetes, Diabetes UK. Available at: https://www.diabetes.org.uk/guide-to-diabetes/complications (Accessed: 04 July 2024).

[11] Klein, H.E. (2023) Diabetes prevalence expected to double globally by 2050, AJMC. Available at: https://www.ajmc.com/view/diabetes-prevalence-expected-to-double-globally-by-2050 (Accessed: 04 July 2024).

[12] Rabbani, N., Kim, G.Y.E., Suarez, C.J. and Chen, J.H. (2022). Applications of machine learning in routine laboratory medicine: Current state and future directions. Clinical Biochemistry. doi:https://doi.org/10.1016/j.clinbiochem.2022.02.011.

[13] Toma, M. and Wei, O.C. (2023) Predictive modeling in medicine, MDPI. Available at: https://www.mdpi.com/2673-8392/3/2/42 (Accessed: 04 July 2024).

[14] Sugandh, F. et al. (2023) Advances in the management of diabetes mellitus: A focus on personalized medicine, Cureus. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10505357/ (Accessed: 05 July 2024).

[15] S, P. (2024) Logistic regression model: A guide to machine learning techniques and applications, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/building-an-end-to-end-logistic-regressionmodel/#:~:text=Logistic%20regression%20is%20a%20Machinelikelihood%20of%20a%20specific%20outcome. (Accessed: 05July 2024).

[16] Olusanya, M.O. et al. (2022) Accuracy of machine learning classification models for the prediction of type 2 diabetes mellitus: A systematic survey and meta-analysis approach, International journal of environmental research and public health. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9655196/ (Accessed: 05 July 2024).

[17] Nguyen B.P., Pham H.N., Tran H., Nghiem N., Nguyen Q.H., Do T.T., Tran C.T., Simpson C.R. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput. Methods Programs Biomed. 2019;182:105055. doi: 10.1016/j.cmpb.2019.105055.

[18] Rajendra, P. and Latifi, S. (2021) Prediction of diabetes using logistic regression and ensemble techniques, Computer Methods and Programs in Biomedicine Update. Available at: https://www.sciencedirect.com/science/article/pii/S26669900210 00318#sec0002 (Accessed: 05 July 2024).

[19] Anderson A.E., Kerr W.T., Thames A., Li T., Xiao J., Cohen M.S. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. J. Biomed. Inform. 2016;60:162–168.doi: 10.1016/j.jbi.2015.12.006.

[20] Kalil A.C., Mattei J., Florescu D.F., Sun J., Kalil R.S. Recommendations for the assessment and reporting of multivariable logistic regression in transplantation literature. Am. J. Transplant. 2010;10:1686–1694. doi: 10.1111/j.1600-6143.2010.03141.x.

[21] Advantages and disadvantages of logistic regression (2024) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/ (Accessed: 05 July 2024).

[22] Joshi, R.D. and Dhakal, C.K. (2021) Predicting type 2 diabetes using logistic regression and machine learning approaches, International journal of environmental research and public health. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8306487/ (Accessed: 05 July 2024).

[23] Sharma, A. (2024) 4 simple ways to split a decision tree in machine learning (updated 2024), Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/ (Accessed: 05 July 2024).

[24] Cerchia, C. and Lavecchia, A. (2023) New Avenues in artificial-intelligence-assisted drug discovery, Science direct. Available at: https://www.sciencedirect.com/science/article/pii/S1359644623000326 (Accessed: 05 July 2024).

[25] Iparraguirre-Villanueva, O. et al. (2023) Application of machine learning models for early detection and accurate classification of type 2 diabetes, U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10378239/ (Accessed: 05 July 2024).

[26] Smolic, H. (2023) Demystifying ai decision trees: A practical guide to understanding and implementing, Graphite Note. Available at: https://graphite-note.com/demystifying-ai-decision-trees-a-practical-guide-to-understanding-and-implementing/#:~:text=However%2C%20like%20any%20AI%20model,poor%20generalization%20on%20unseen%20data (Accessed: 05 July 2024).

[27] Random Forest (no date) Random Forest - an overview | ScienceDirect Topics. Available at: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/random-forest#:~:text=It%20is%20based%20on%20the,of%20the%20data%20and%20features. (Accessed: 05 July 2024).

[28] Sharma, A. (2024) Random Forest vs decision tree: Which is right for you?, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ (Accessed: 05 July 2024).

[29] Wang, S. (2024) Diabetes prediction using random forest in Healthcare, Highlights in Science, Engineering and Technology. Available at: https://drpress.org/ojs/index.php/HSET/article/view/19875 (Accessed: 05 July 2024).

[30] Mahboob Alam , T. et al. (2019) A model for early prediction of diabetes, Informatics in Medicine Unlocked. Available at: https://www.sciencedirect.com/science/article/pii/S2352914819300176 (Accessed: 05 July 2024).

[31] Ahmed, N. et al. (2021) Machine learning based diabetes prediction and development of Smart Web Application, International Journal of Cognitive Computing in Engineering. Available at: https://www.sciencedirect.com/science/article/pii/S2666307421000279#:~:text=5.4.&text=For%20NB%2C%20DT%2C%20RF%2C,and%20exceed%20the%20other%20approaches (Accessed: 05 July 2024).

[32] Couronné, R., Probst, P. and Boulesteix, A.-L. (2018) Random Forest versus logistic regression: A large-scale benchmark experiment - BMC Bioinformatics, BioMed Central. Available at: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5 (Accessed: 05 July 2024).

[33] Sharma, T. and Shah, M. (2021) A comprehensive review of machine learning techniques on diabetes detection, Visual computing for industry, biomedicine, and art. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8642577/ (Accessed: 05 July 2024).

[34] Aishwarya Mujumdar (2020) Diabetes prediction using machine learning algorithms, Procedia Computer Science. Available at: https://www.sciencedirect.com/science/article/pii/S1877050920300557?via%3Dihub (Accessed: 05 July 2024).

[35] GeeksforGeeks (2023) Support Vector Machine in machine learning. Available at: https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/ (Accessed: 05 July 2024).

[36] Alehegn, M., Joshi, R. and Alehegn, M., (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, Volume no 4 , Issue no 10, page no 426-436.

[37] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H., (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, Volume no 9, page no 115.

[38] Aslan, M.F. and Sabanci, K. (2023). A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data. Diagnostics, 13(4), p.796. doi:https://doi.org/10.3390/diagnostics13040796.

[39] Swapna G, Soman Kp, Vinayakumar R. (2018). Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Computer Science, 132, pp.1253–1262. doi:https://doi.org/10.1016/j.procs.2018.05.041.

[40] Muthu, Joanish & Suriya, s. (2023). Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm. Journal of Trends in Computer Science and Smart Technology. 5. 10.36548/jtcsst.2023.2.007.

[41] Srivastava, T. (2019). Introduction to KNN, K-Nearest Neighbors : Simplified. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/.

[42] Thakur, D., Gera, T., Bhardwaj, V., Ahmad Ali AlZubi, Ali, F. and Singh, J. (2023). An enhanced diabetes prediction amidst COVID-19 using ensemble models. Frontiers in Public Health, 11. doi:https://doi.org/10.3389/fpubh.2023.1331517.

[43] Edeh M.O., Khalaf O.I., Tavera C.A., Tayeb S., Ghouali S., Abdulsahib G.M., Richard-Nnabu N.E., Louni A. (2022) A Classification Algorithm-Based Hybrid Diabetes Prediction Model. Front. Public Health. 10:829519. doi: 10.3389/fpubh.2022.829519.

[44] Chen W., Chen S., Zhang H., Wu T. (2017) A hybrid prediction model for type 2 diabetes using K-means and decision tree; Proceedings of the 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS); Beijing, China. 24–26; pp. 386–390.

[45] Jadama, Ansumana & Toray, Modou. (2024). Ensemble Learning: Methods, Techniques, Application. 10.13140/RG.2.2.28017.08802.

[46] Turke Althobaiti, Saad Althobaiti and Selim, M.M. (2024). An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making. Alexandria Engineering Journal /Alexandria Engineering Journal, 94, pp.311–324. doi:https://doi.org/10.1016/j.aej.2024.03.044.

[47] Mustafa, M. (2023) Diabetes prediction dataset, Kaggle. Available at: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data (Accessed: 05 July 2024).

[48] How do you preprocess data and engineer features in your machine learning model deployment? (2024) Data Preprocessing and Feature Engineering for ML Model Deployment. Available at: https://www.linkedin.com/advice/0/how-do-you-preprocess-data-engineer-features-jkhgc#:~:text=Data%20preprocessing%20and%20feature%20engineering%20can%20bring%20many%20benefits%20to,errors%2C%20and%20facilitating%20debugging%20and (Accessed: 05 July 2024).

[49] Ray, S. (2022) Simple methods to deal with categorical variables in predictive modeling, Analytics Vidhya. Available at: https://analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/ (Accessed: 05 July 2024).

[50] Ehrenstein, V., Kharrazi, H., Lehmann, H. and Taylor, C.O. (2019). Obtaining Data From Electronic Health Records. [online] www.ncbi.nlm.nih.gov. Agency for Healthcare Research and Quality (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK551878/.

[51] Zaher, M., Ghoneim, A.S., Abdelhamid, L. and Atia, A. (2024). Unlocking the potential of RNN and CNN models for accurate rehabilitation exercise classification on multi-datasets. Multimedia tools and applications. doi:https://doi.org/10.1007/s11042-024-19092-0.

[52] Abeysinghe, A., Tohmuang, S., Davy, J.L. and Fard, M. (2023). Data augmentation on convolutional neural networks to classify mechanical noise. Applied Acoustics, 203, p.109209. doi:https://doi.org/10.1016/j.apacoust.2023.109209.

[53] Aslam, Nida & Khan, Irfan & Alkhalifah, Samar & AL-Sadiq, Sarah & Bughararah, Shahad & AL-Otabi, Meznah & AL-Odinie, Zainab. (2021). Predicting Diabetic Patient Hospital Readmission Using Optimized Random Forest and Firefly Evolutionary Algorithm. International Journal on Advanced Science, Engineering and Information Technology. 11. 1876. 10.18517/ijaseit.11.5.14221.

[54] Mohsen, F., Al-Absi, H.R.H., Yousri, N.A., El Hajj, N. and Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. npj Digital Medicine, [online] 6(1), pp.1–15. doi:https://doi.org/10.1038/s41746-023-00933-5.

[55] Khan, A.A., Chaudhari, O. and Chandra, R. (2023). A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced problems: Combination, Implementation and Evaluation. Expert Systems with Applications, [online] 244, p.122778. doi:https://doi.org/10.1016/j.eswa.2023.122778.

[56] Krichen, Moez. (2023). Convolutional Neural Networks: A Survey. Computers. 12. 151. 10.3390/computers12080151.

[57] Azam, M., Sadman Sakib, Nur Mohammad Fahad, Abdullah Al Mamun, Md. Anisur Rahman, Swakkhar Shatabda and Hossain, S. (2024). A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks. Decision analytics journal, pp.100470–100470. doi:https://doi.org/10.1016/j.dajour.2024.100470.

[58] Singh, P. (2023). Systematic review of data-centric approaches in artificial intelligence and machine learning. doi:https://doi.org/10.1016/j.dsm.2023.06.001.

[59] Grote, T. and Keeling, G. (2022) Enabling fairness in healthcare through Machine Learning - Ethics and Information Technology, SpringerLink. Available at: https://link.springer.com/article/10.1007/s10676-022-09658-7 (Accessed: 05 July 2024).

[60] Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A., Almohareb, S.N., Aldairem, A., Alrashed, M., Saleh, K.B., Badreldin, H.A., Yami, A., Harbi, S.A. and Albekairy, A.M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education, [online] 23(1). doi:https://doi.org/10.1186/s12909-023-04698-z.

[61] Nass, S.J. (1970) The value and importance of Health Information Privacy, Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Available at: https://www.ncbi.nlm.nih.gov/books/NBK9579/ (Accessed: 05 July 2024).

[62] What is Data Security? (2024) DataGuard. Available at: https://www.dataguard.co.uk/blog/what-is-data-security/#:~:text=By%20adhering%20to%20regulations%20like,to%20brand%20reputation%2C%20and%20lawsuits. (Accessed:05 July 2024).

[63] McKendrick , J. and Thurai, A. (2022) Ai isn’t ready to make unsupervised decisions, Harvard Business Review. Available at: https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions (Accessed: 05 July 2024).

[64] Osasona, Femi & Amoo, Olukunle & Atadoga, Akoh & Abrahams, Temitayo & Farayola, Oluwatoyin & Ayinla, Benjamin. (2024). REVIEWING THE ETHICAL IMPLICATIONS OF AI IN DECISION MAKING PROCESSES. International Journal of Management & Entrepreneurship Research. 6. 322-335. 10.51594/ijmer.v6i2.773.