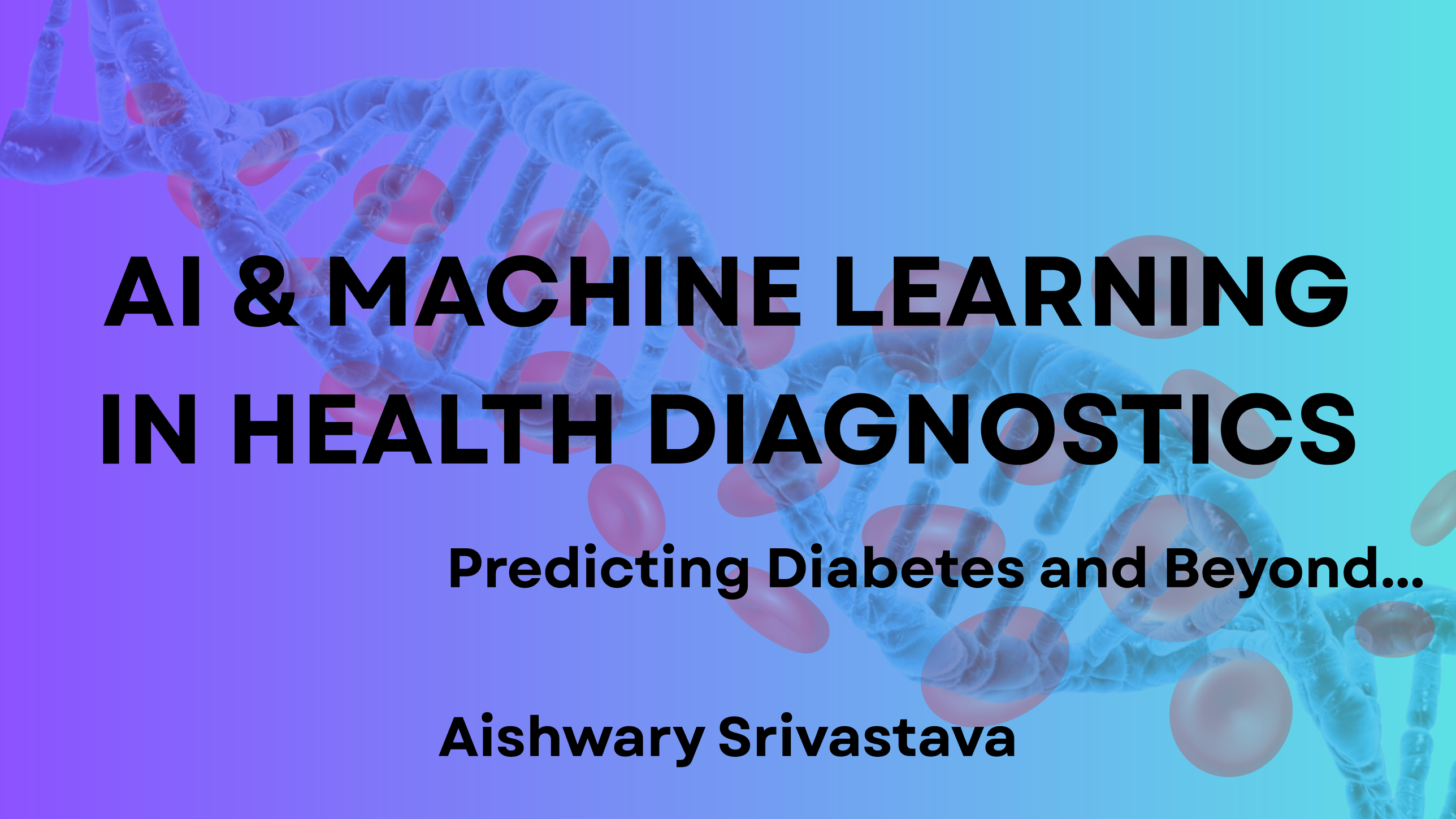


AI & MACHINE LEARNING IN HEALTH DIAGNOSTICS

Predicting Diabetes and Beyond...

Aishwary Srivastava

INTRODUCTION

Diabetes is a chronic condition that affects millions of people worldwide. India is the diabetes capital of the world, where one in every ten people is diabetic. It is said that more than half of diabetic cases go undiagnosed in India therefore early diagnosis and prediction of diabetes can help in timely medical intervention.

DATASET OVERVIEW

The dataset used for this project contains 100,000 patient records with 9 key health indicators relevant to diabetes prediction. These features help AI models assess a patient's risk based on routine health parameters.

Health Indicators :

- Gender
- Age
- Hypertension
- Heart Disease
- Smoking History
- BMI
- HbA1c Level
- Blood Glucose Level
- Diabetes

DATA PREPROCESSING

- **Missing Values Handling – Ensuring no gaps in the data.**
- **Feature Encoding – Converting categorical data (e.g., Gender, Smoking History) into numerical format.**
- **Standardization – Scaling numerical features (Age, BMI, HbA1c, Blood Glucose) for better model performance.**

MACHINE LEARNING MODELS

- Logistic Regression
- Support Vector Machine (SVM)
- Neural Networks (NN)

Logistic Regression :

- Logistic Regression is a statistical method used for binary classification problems. It estimates the probability of an instance belonging to a particular class using a logistic function.

- Our logistic regression model utilises the linear solver (Liblinear), which is suitable for relatively small datasets and provides good classification performance.

Support Vector Machine (SVM):

- SVM is a machine learning algorithm that classifies patients as diabetic or non-diabetic by finding the best hyperplane to separate the two groups. It works by maximizing the margin between classes, making it highly effective for structured medical data

- We used the Radial Basis Function (RBF) kernel to capture complex relationships between features like BMI, blood glucose levels, and smoking history.

Neural Networks (NN) :

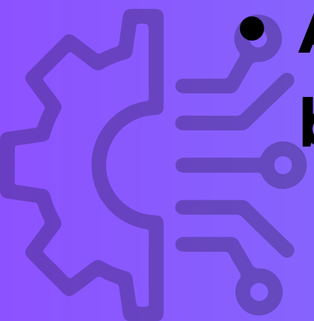
- Our deep learning framework utilizes two hidden layers where ReLU acts as the activation function and an output layer that utilizes sigmoid as the activation function to give a binary final output (0 for No diabetes and 1 for Diabetes).

- This architecture allows the model to learn complex patterns and relationships within the health data that simpler models might miss.

MODEL EVALUATION

Evaluation Metrics :

- Accuracy – Overall correctness of predictions
- Precision – Correct positive identifications.
- Recall – Ability to detect actual diabetic patients.
- F1 Score – Balance between precision & recall.
- ROC Curve – Graphical representation of model performance.
- AUC Score – Measures how well the model distinguishes between classes.

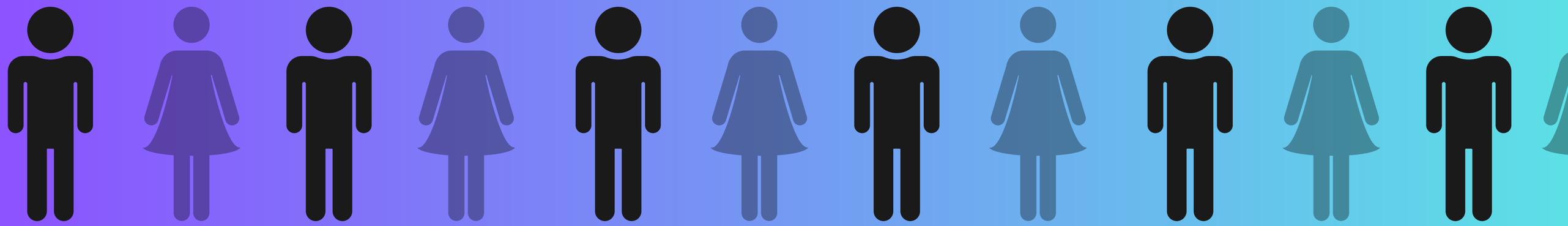


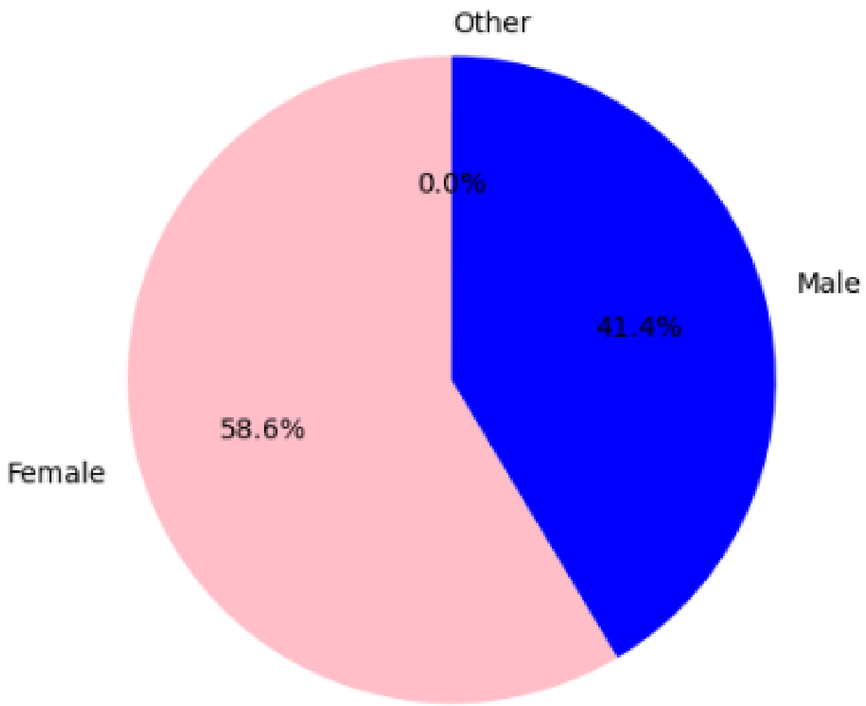
COMPARISON OF MODELS :

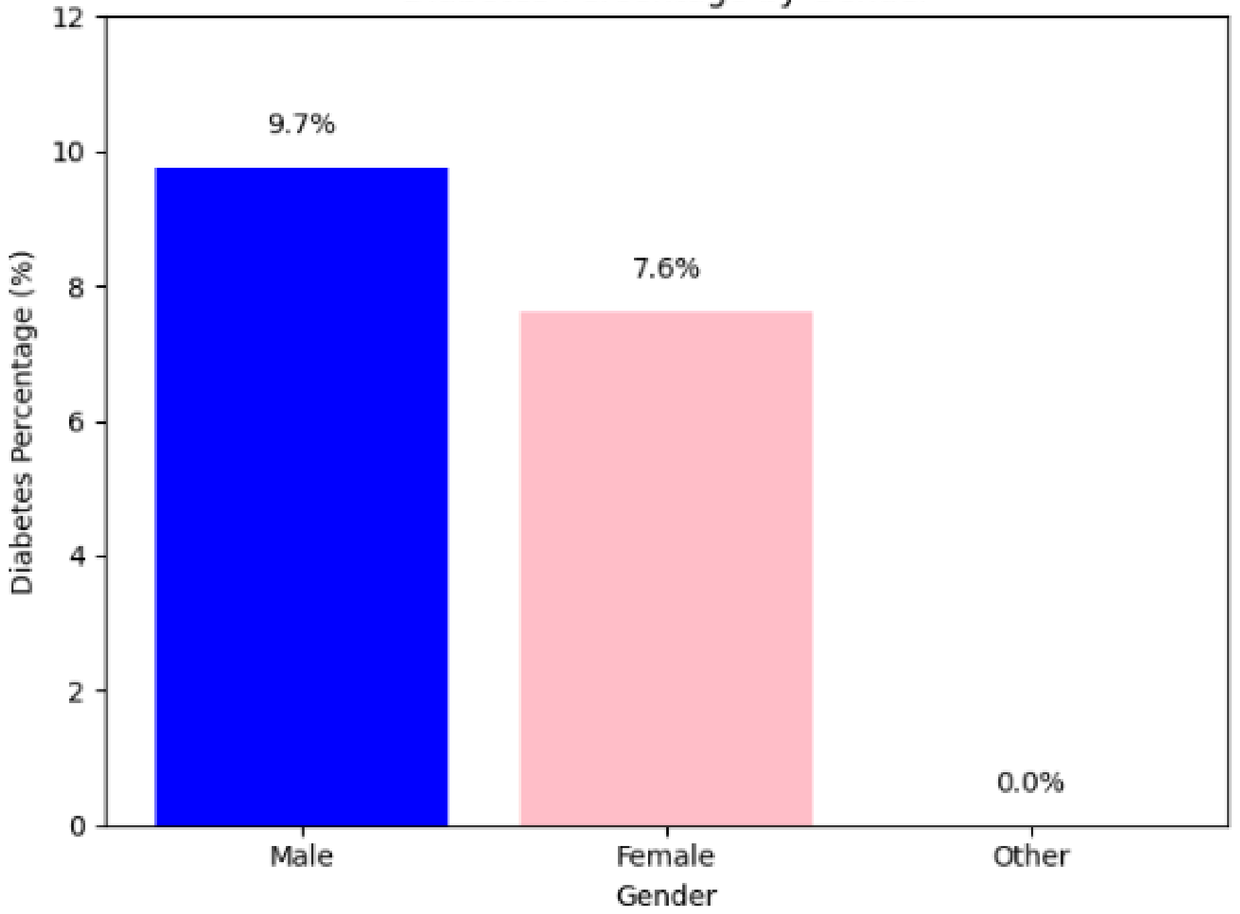
Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-score (Class 1)	ROC AUC
Logistic Regression	96.07%	86.38%	63.86%	73.43%	0.9621
SVM	96.34%	96.75%	58.87%	73.20%	0.9052
Deep Learning	**97.23%**	**97.99%**	**68.80%**	**80.84%**	**0.9760**

KEY FINDINGS

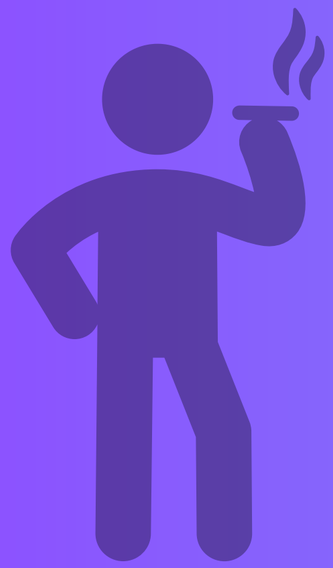
- **According to the dataset, diabetes is more prevalent in Males than Females. This gender disparity could be attributed to various lifestyle and physiological factors.**

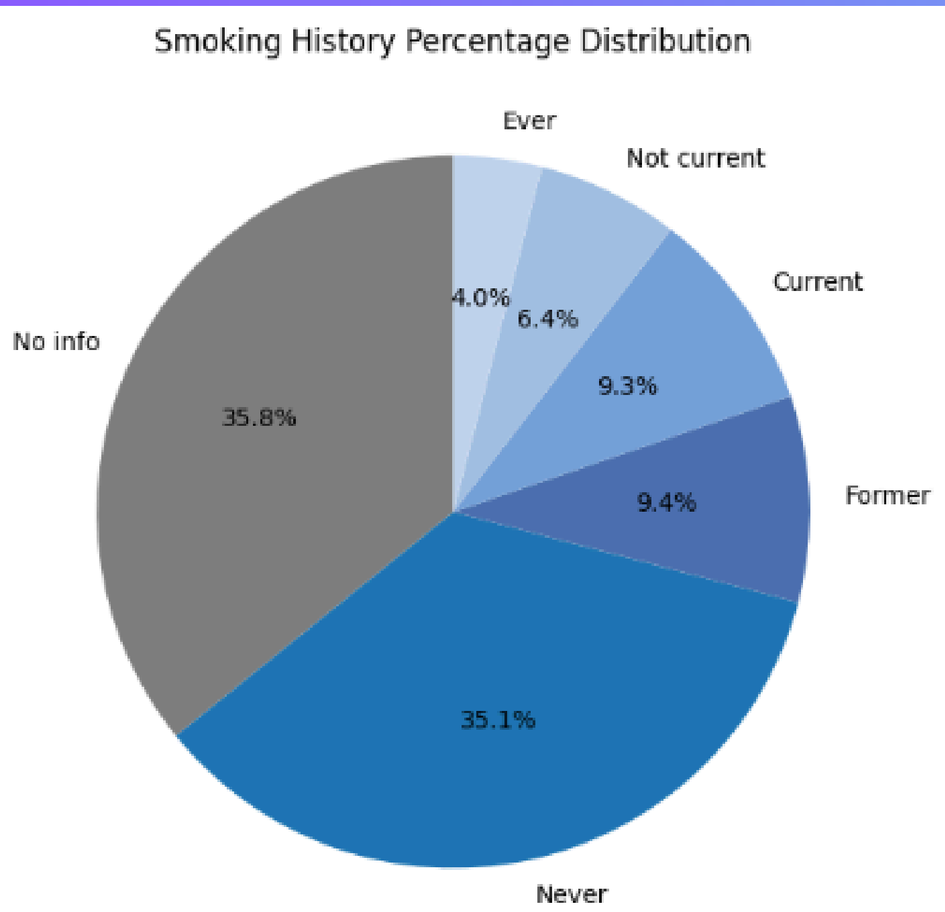


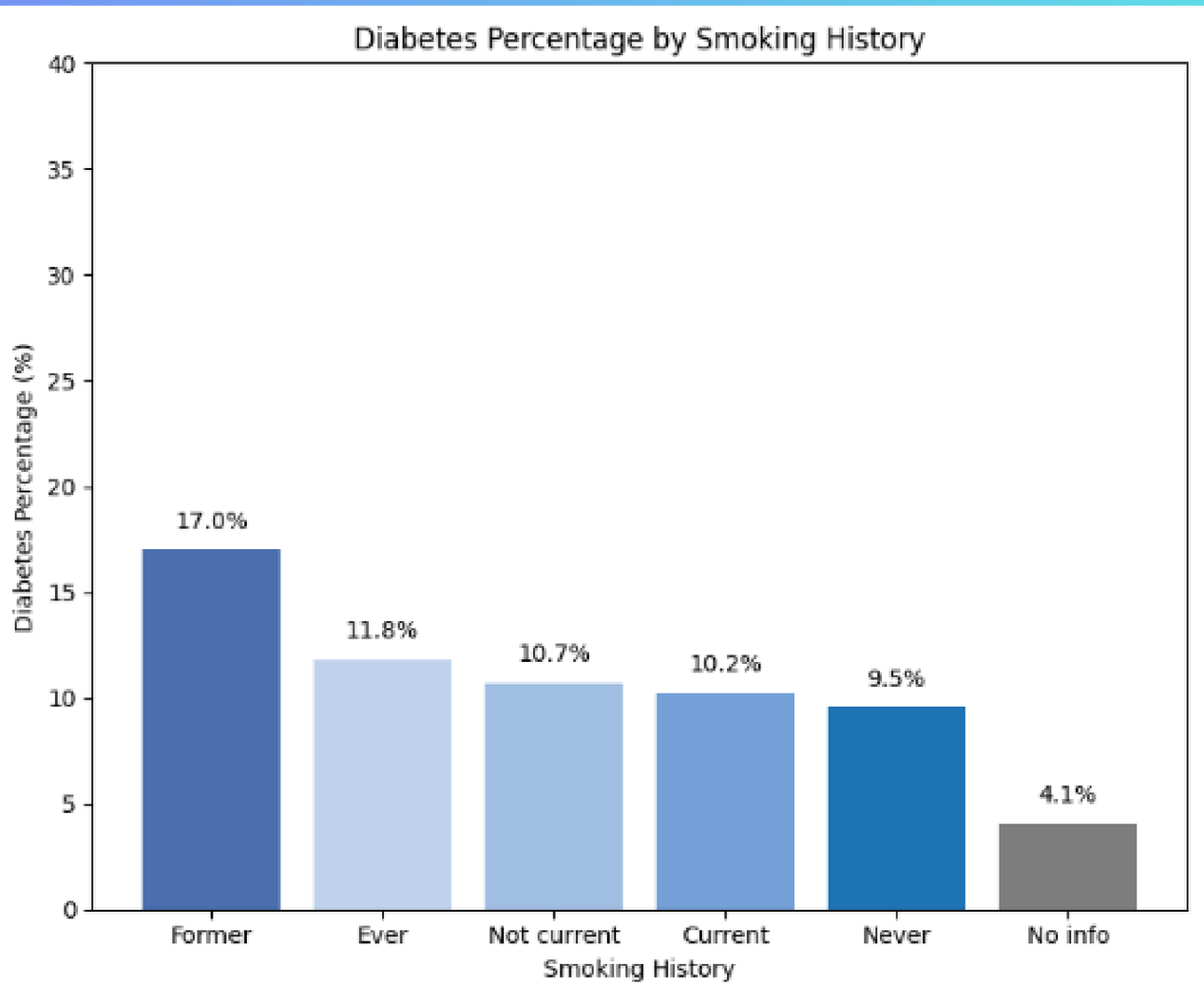

Gender Percentage Distribution in the dataset
Other
0.0%
Male
41.4%
58.6%
Female


Diabetes Percentage by Gender
Diabetes Percentage (%)
12
10
8
6
4
2
0
9.7%
7.6%
0.0%
Male
Female
Other
Gender

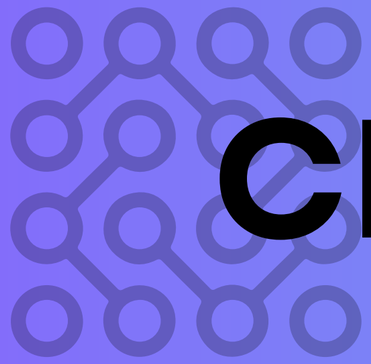
- **Smokers, in general, are more prone to having diabetes than non-smokers. We observed a clear correlation between smoking history and diabetes risk, with former smokers showing the highest risk, followed by patients who have smoked since forever.**




Smoking History Percentage Distribution
Ever
Not current
Current
Former
Never
No info
4.0%
6.4%
9.3%
9.4%
35.1%
35.8%


Diabetes Percentage by Smoking History
Diabetes Percentage (%)
Smoking History
Former
Ever
Not current
Current
Never
No info
17.0%
11.8%
10.7%
10.2%
9.5%
4.1%
0
5
10
15
20
25
30
35
40

- **HbA1c level and Blood Glucose Level were identified as the strongest predictors of diabetes, as expected, given their direct relationship with blood sugar regulation.**



CHALLENGES

- The first challenge was to decide on the encoding strategy that was to be used on the smoking history column since the values were diverse yet impactful on the final output.

- The second challenge was to manually tune the hyperparameters of the models and decide what works best.

FUTURE IMPROVEMENTS

- Hyperparameter Tuning
- Feature Selection
- Training on more Data
- Model Deployment

CONCLUSION

AI-powered diagnostics can revolutionize early disease detection, improving patient outcomes. Our Neural Networks model achieved the highest accuracy, effectively predicting diabetes risk using key health indicators. With further refinements, AI can enhance preventive care and personalized treatment, making healthcare more efficient and accessible.