Final Project Report 171

Laurel Jasper Murphy*, Gezheng Kang†, Ariana Smith‡, Ayushi Kishore§, Varsha Sivaprakash¶

*Dept. of Computer Science, UC Davis, Davis, CA, USA, ljamurphy@ucdavis.edu

†Dept. of Computer Science and Engineering, UC Davis, Davis, CA, USA, gzkang@ucdavis.edu

‡Dept. of Computer Science, UC Davis, Davis, CA, USA, masmith@ucdavis.edu

§Dept. of Computer Science, UC Davis, Davis, CA, USA, aykishore@ucdavis.edu

¶Dept. of Computer Science, UC Davis, Davis, CA, USA, varsivaprakash@ucdavis.edu

Abstract—This study uses data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) to develop a model that predicts diabetes status, categorized as no diabetes or diabetes. The target variable is based on self-reported medical diagnosis, excluding cases related solely to pregnancy. The dataset includes 21 predictive features capturing cardiovascular risk factors (e.g., high blood pressure, high cholesterol), behavioral indicators (e.g., smoking history, alcohol consumption, physical activity), demographic attributes (e.g., age, sex, education, household income), and general health conditions (e.g., obesity, sleep, and mental health). These features collectively provide a comprehensive overview of factors contributing to diabetes risk. We utilized and tested many models.

We evaluated multiple models, including logistic regression, Bayesian networks, random forest, XGBoost, LightGBM, and several neural networks. Initial models achieved prediction accuracies ranging from 75% to 88%, but suffered from poor generalization due to suboptimal feature selection. After reworking the feature set and refining the model architecture, we observed improvements across all models, with accuracies exceeding 90 percent. We then scaled our neural network to an 11-layer architecture with 30 neurons per layer and ReLU activation. This final model achieved a test accuracy of 91.01%. Although it did not achieve the highest absolute accuracy among all tested models, it demonstrated the strongest generalization performance across key evaluation metrics.

Index Terms—diabetes prediction, Behavioral Risk Factor Surveillance System, machine learning, logistic regression, neural networks, public health

I. INTRODUCTION AND BACKGROUND

Diabetes is a chronic metabolic disorder characterized by prolonged elevated blood glucose levels. This condition arises either from insufficient insulin production by the pancreas or the body's inability to effectively respond to insulin [1]. The disease disrupts carbohydrate, fat, and protein metabolism and can lead to serious health complications, including cardiovascular disease, kidney failure, neuropathy, and vision loss. In the past few decades, diabetes has transformed from a rare disorder into a widespread public health crisis. The World Health Organization now classifies diabetes as one of the top ten causes of death worldwide, with projections indicating continuous growth in prevalence.

The human cost of diabetes is severe. Research indicates that individuals with diabetes are approximately 1.8 times more likely to die from any cause and over 2.3 times more likely to die from vascular disease compared to those without the condition. In terms of life expectancy, a person of 50 years with diabetes loses, on average, six years of life. The disease also increases susceptibility to various cancers, including those of the pancreas, liver, breast, and colon. The longer a person lives with undiagnosed or uncontrolled diabetes, the more likely they are to suffer irreversible health consequences. [2] These facts underscore the critical importance of early detection and proactive management.

From an economic perspective, the burden is equally significant. According to estimates, the global cost of diabetes surpasses USD 1.3 trillion annually, accounting for nearly 1.8 percent of the global GDP. These expenses encompass both direct healthcare costs (such as hospitalizations, medications, and physician visits) and indirect costs (including lost productivity, premature mortality, and long-term disability). Particularly in low and middle income countries, where healthcare access is limited, the indirect costs are disproportionately high, leading to further socioeconomic disparities in outcomes. [3] Without significant improvements in prevention, early screening, and chronic disease management, these costs are expected to escalate dramatically in the coming decades.

To address the growing crisis of the diabetes pandemic, new technologies must be explored to assist with early detection and intervention. One of the most promising frontiers is artificial intelligence (AI) and machine learning (ML), which can process complex, high-dimensional datasets and identify subtle patterns that traditional methods may overlook. These technologies offer the potential to provide inexpensive, scalable, and non-invasive diagnostic support, particularly in underserved areas where medical personnel or equipment may be scarce.

This project examines the intersection of AI and healthcare by developing a predictive model for diabetes risk. Using data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), we leverage a cleaned and balanced version of the Diabetes Health Indicators dataset. This dataset comprises 21 behavioral and demographic features, including physical health, diet, body mass index (BMI), education, income, and access to healthcare services. Rather than treating diabetes as a three-class problem (type 0, 1, or 2), we simplified the task into a binary classification—predicting whether an individual does or does not have diabetes.

Our motivation for choosing this topic stems from both the scale of the public health crisis and the opportunity to create a meaningful, real-world application of AI. Early diagnosis of diabetes has been shown to reduce complications and improve quality of life significantly. However, current screening methods are often reactive rather than proactive, missing opportunities to identify individuals at risk before symptoms manifest. By building a machine learning model trained on behavioral and self-reported data, we aim to develop a method for early detection that can be deployed in mobile applications, telehealth platforms, or community screening initiatives. We tested several modeling approaches, including logistic regression and Bayesian classifiers, but ultimately focused on designing a deep learning model: an eleven-layer artificial neural network. This network was optimized through grid search, hyperparameter tuning, and extensive evaluation, achieving around 91 percent accuracy on held-out data. While this accuracy reflects room for improvement, our research highlights the feasibility of developing an AI-powered screening tool that strikes a balance between precision, recall, and interpretability.

This work contributes to a growing body of research on AI in healthcare. It demonstrates how predictive models can be trained on accessible datasets to assist in early disease detection. In the future, such models could be integrated into broader preventive healthcare strategies, serving as early warning systems that alert individuals and healthcare providers to elevated risk—before disease progression.

II. Literature Review

Recent literature emphasizes the importance of selecting appropriate modeling techniques and preprocessing strategies for health-related classification tasks. Traditional models like binomial logistic regression remain effective for categorical datasets, particularly when regularization and optimal feature selection are applied, as shown by high accuracy and low AIC scores. Neural networks, on the other hand, are well-suited to capturing complex, non-linear patterns in one-hot encoded categorical data, with architectures benefiting from techniques such as L1/L2 regularization, ReLU activation, and hyperparameter tuning via tools like KerasTuner. Overall, such approaches demonstrate the evolving landscape of machine learning in medical diagnostics, where model performance hinges on careful data transformation, architecture design, and context-specific evaluation metrics.

A robust body of research supports the inclusion of the 21 features in our diabetes prediction model. Clinical studies confirm that high blood pressure, BMI, and smoking are strong predictors of diabetes due to their links with insulin resistance and inflammation ([4]; [5]; [6]). Dietary habits, such as low fruit and vegetable intake, also correlate with increased diabetes risk, while regular physical activity offers protective benefits [7]; [8]; [9]). Cholesterol levels and regular cholesterol checks are relevant due to evidence of low HDL in some diabetes subtypes ([10]). Mental health and general health status are important, as diabetes often coexists with depression and chronic illness ([11]). Socioeconomic factors—including income, education, insurance status, and healthcare access—influence diagnosis and management, particularly in underserved communities ([12]; [13]; [14]; [15]). Cost-related barriers often delay care, worsening outcomes (IHPI). Demographics like age and sex affect susceptibility, with older adults and postmenopausal women at higher risk ([16]; [17]). Finally, difficulty walking is linked to diabetes-related neuropathy and vascular dysfunction ([18]). Together, these features enhance the accuracy and real-world applicability of our predictive model.

III. Dataset Description and Exploratory Data Analysis

Dataset description: The dataset we used was the Diabetes Health Indicators from the CDC. This dataset consisted of 3 classes of diabetes and 21 features. 0 indicated no diabetes, 1 indicated type 1 diabetes and 2 indicated type 2 diabetes. We treated this problem as a binary classification task, where the predictions indicate whether or not a person has diabetes based on a given set of features. Upon observing the distribution of the features and diabetes classes, we quickly realized that there was a huge class imbalance between positive and negative cases of diabetes, with over 80 percent of the dataset having no diabetes.
To mitigate this imbalance, we found the same dataset, so that the ratio between the positive and negative

diabetes cases have a 50:50 ratio. Here, the negative cases were indicated by 0 and positives were indicated by 1, while still having 21 features. Some of the features were continuous values, while most were categorical values. Below are the features:

- **Continuous:** MentHlth, PhyHlth, BMI
- **Discrete:** HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, DiffWalk, Sex, Age, Education, Income
- **Target (Discrete 0 or 1):** `Diabetes_binary`

Exploratory Data Analysis:

Before any data cleaning, all model accuracies were roughly around 75%. We computed the linear correlation using pearson correlation between continuous and continuous pairs and point-biserial correlation between categorical and continuous pairs. Based on these computations, none of the feature pairs had strong linear correlation. The first attempt of cleaning involved outlier detection and removal. Using the interquartile range method, we were able to detect outliers for all continuous variables: BMI, Mental Health and Physical Health. We decided to remove these outliers to avoid the skewed data from disrupting our model performances. However, we noticed how the data points were almost halved upon removing the outliers, indicating that the outliers must be significant to the model's predictions.

Next, we conducted a literature review for all 21 features which confirmed the correlation of all features with diabetes, including the continuous features for which the outliers were removed.

To reduce the effects of the skewed distribution of the continuous variables, we decided to categorize them into 4-5 classes each based on their value ranges. This change was the key part of our EDA, after which the model accuracies significantly improved, improved to approximately 90% accuracy. Upon further plotting, the data was significantly less skewed compared to pre-categorization.

Finally, on our now fully categorical dataset, we decided to identify the most important features using the Chi-Squared coefficient which is used for detecting non-linear dependencies between categorical variables. Based on this, the following features were selected as the most important:

- `HighBP`
- `HighChol`
- `NEW_BMI_class`
- `Stroke`
- `HeartDiseaseorAttack`
- `PhysActivity`

- `HvyAlcoholConsump`
- `GenHlth`
- `MetHlth_Class`
- `PhyHlth_Class`
- `DiffWalk`
- `Age`
- `Income`

However, most model performances decreased slightly overall when any of the features were removed, indicating that all features carried some importance towards the prediction of diabetes, as supported by our feature literature review. Therefore, we kept all 21 features most of the time during our study unless necessary to remove, since all features have some significance to the prediction of diabetes.

IV. PROPOSED METHODOLOGY

All features in this dataset are categorical, meaning they represent qualitative labels, not numerical quantities. Therefore, standardization is not appropriate. For instance, the HighBP feature uses 0 to indicate no high blood pressure and 1 to indicate presence of high blood pressure. Standardizing this would incorrectly imply that $1 > 0$ in a numerical sense, which distorts the categorical nature of the data. To handle categorical features correctly, we applied one-hot encoding.

In addition, we bucketed the continuous features BMI, MentHlth, and PhysHlth into discrete classes based on meaningful thresholds. For example, BMI is categorized into five classes (0 through 4). More details on the binning criteria can be found in the final text cell of the notebook. Since artificial neural networks (ANNs) are capable of capturing both linear and non-linear relationships, we fed the entire encoded dataset into the model without additional feature engineering.

The model consists of 11 total layers: 1 input layer, 10 hidden layers, and 1 output layer. Each hidden layer contains 30 neurons and uses the ReLU activation function. The output layer uses a sigmoid activation function, suitable for binary classification. L1 and L2 regularization (each set to 0.001) was applied across all layers to reduce overfitting. The SGD optimizer was used with a momentum of 0.9 to accelerate convergence. The ANN achieved high accuracy on the test set, indicating good generalization. The confusion matrix and derived metrics (precision, recall, F1-score) also showed strong performance, confirming that the model effectively captured the relationship between the features and the target variable. The ANN model performed well with the fully one-hot

encoded dataset, validating that proper encoding and a well-structured neural network can effectively model complex relationships in categorical health-related data.

To better understand the learning behavior of **ANN Model A**, we visualized the model's training dynamics over 100 epochs.
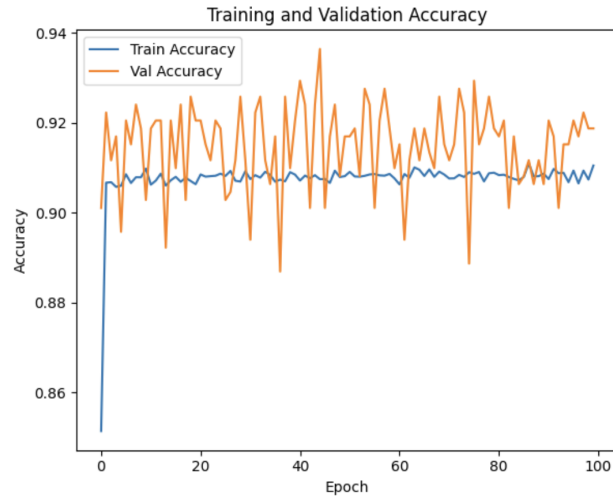

Fig. 1: shows a smooth convergence trend, with both training and validation loss decreasing steadily. Slight overfitting is observed in later epochs, but regularization mitigates major divergence.
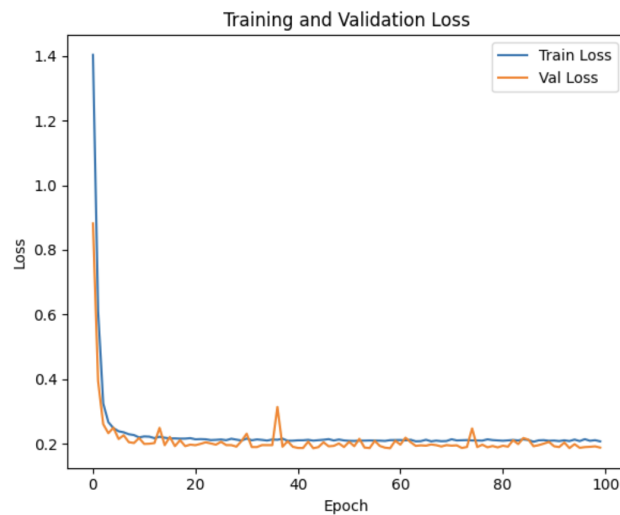

Fig. 2: Indicates consistent accuracy growth for both training and validation sets. The final accuracy plateau suggests successful training convergence.

These trends confirm the model's learning capacity and generalization ability, though fine-tuning can further improve its performance and reduce training time.

V. EXPERIMENTAL RESULTS AND EVALUATION

This section presents the results of our comparative evaluation of eight different classification models, including Logistic Regression, several Artificial Neural Network (ANN) architectures, a Bayesian Network, and tree-based models such as Random Forest, XGBoost, and LightGBM. Each model was evaluated using consistent performance metrics: confusion matrix, accuracy, precision, recall, F1 score, and testing time.

These represent all of the models we trained and tested during this project. The best-performing model, ANN A, was introduced in the previous section.

Binomial logistic regression:
Binomial logistic regression was applied to a fully categorical dataset (3 to classes) with no need for standardization, binary class labels, and equal class weighting. No outliers were removed prior to modeling. The same optimization techniques as in previous models were used, and four models were tested in total. The best-performing models (a and b) used all 21 features, with Model a employing L2 regularization and the 'lbfgs' solver. Model a achieved the highest accuracy (91.06 percent) and a low AIC (5123.826), while Model b had a nearly identical classification report with a slightly lower AIC (5113.101). Models c and d, which used only 13 features, matched Model a's accuracy but had higher AIC scores and the highest number of false negatives (753), a critical factor in medical data contexts. These results indicate that including all 21 features was necessary for optimal classification performance.

ANN A: We designed a deep feedforward neural network with 10 layers, each containing 30 neurons, to classify diabetes risk. The hidden layers employed the ReLU activation function to introduce non-linearity and enable the network to learn complex patterns, while the output layer used a sigmoid activation for binary classification. To prevent overfitting, we applied both L1 and L2 regularization with a weight of 0.001. The model was trained using Stochastic Gradient Descent (SGD) enhanced with a momentum of 0.9 to accelerate convergence and avoid local minima. We used a learning rate of 0.01, trained the network for 100 epochs, and set the batch size to 64. The loss function was binary cross-entropy, which is suitable for our binary classification task. This configuration aimed to leverage depth and regularization to achieve a well-generalized model for predicting diabetes risk.

ANN B:
After exploratory data analysis, we implemented a binary classifier for diabetes risk. We split the data 80-20 into train and test sets, and constructed a six-layer feedforward Artificial Neural Network with Tensorflow/Keras. The hidden layers utilized the ReLU activation function, and we tried different L1 and L2 regularization weights to prevent overfitting of the model. The output layer uses the default sigmoid function and binary cross-entropy loss to improve the binary classifier. We implemented a brute-force grid search with five hyperparameters to hyperparameter tune the ANN, an SGD learning rate and momentum, a hidden layer with either 16 or 32 neurons, and L1/L2 regularization. We also ran this over 50 epochs and a batch size of 64. This brute force grid search method generalized our eleven-layer network's hyperparameters and provided a reproducible model from the root BRFSS features to a tuned diabetes-risk predictor.

ANN T and KerasTuner:
Implemented a supervised machine learning pipeline to predict diabetes using a feedforward neural network with hyperparameter tuning. The process begins with loading and preprocessing a dataset, where the target variable is extracted and categorical features are one-hot encoded. The data is then split into training and testing sets, followed by standardization using StandardScalar. A custom model-building function is defined using KerasTuner, allowing for dynamic architecture adjustments, including varying the number of layers, units per layer, and learning rate. A RandomSearch tuner is employed to identify the best model configuration by optimizing for validation accuracy. Early stopping is incorporated to prevent overfitting. The best model is re-trained and evaluated on the test set, reporting both accuracy and optimal hyperparameters. Finally, the training history, confusion matrix, and other parameters are visualized to show trends in loss and accuracy over epochs, providing insight into the model's learning behavior and results.

Bayesian Network (BN):
To model probabilistic dependencies among features and predict diabetes risk, we constructed a Bayesian Network using the Hill Climb Search algorithm for structure learning. The K2 scoring metric guided the search for the optimal network topology, balancing model complexity and data fit. We applied Variable Elimination for efficient inference, allowing us to compute conditional probabilities and make probabilistic predictions. To simplify the model and improve generalizability, we reduced the feature space to 13 key attributes based on prior analysis.

This approach enabled interpretable reasoning under uncertainty while maintaining computational tractability.

Random Forest:
We implemented Random Forest using the Random Forest Classifier from scikit-learn. All features were one-hot encoded to handle categorical variables properly. To address the class imbalance in the target variable, we used class_weight='balanced', which automatically adjusts the weight inversely proportional to class frequencies. The model was trained and validated using 5-fold stratified cross-validation, and performance was evaluated using accuracy, precision, recall, F1 score, and testing time. Finally, the model was retrained on the full training set and tested on the held-out test set to assess generalization.

XGBoost:
The XGBoost model was implemented using XGBClassifier. We configured the objective as 'binary logistic' for binary classification and used 'logloss' as the evaluation metric. The tree_method='hist' was selected for faster training on structured data. Because XGBoost does not natively support class_weight, we manually computed sample weights based on the class distribution and passed them during model fitting. Similar to Random Forest, we performed 5-fold stratified cross-validation to evaluate the model, followed by a final evaluation on the test set. The model's robustness and fast inference make it attractive for structured medical data.

LightGBM:
We used the LGBMClassifier to implement the Light-GBM model, setting the objective to 'binary'. The parameter class_weight='balanced' ensured that the model addressed class imbalance effectively. We used 100 estimators and a fixed random_state=42 for reproducibility. LightGBM's ability to handle large numbers of features and fast inference made it particularly suitable for our one-hot encoded dataset. As with the other models, the evaluation was conducted by five-fold cross-validation, followed by a final evaluation of the test set to confirm the performance and generalization of the model.

Performance Report:
To better understand the performance and practical suitability of each model, we conducted a comprehensive evaluation across both quantitative and qualitative criteria. This involved analyzing core metrics such as precision, recall, and F1 score, alongside factors like interpretability, scalability, and inference efficiency. The following figure summarizes our findings.

Model	Key Hyperparameters
Logistic Regression (L2)	Solver: 'lbfgs', Penalty: ℓ_2, C = 1.0, AIC = 5123.826
Logistic Regression (L1)	Solver: 'liblinear', Penalty: ℓ_1, C = 1.0, AIC = 5113.101
ANN Model A	10 layers × 30 neurons, ReLU, Sigmoid, L1L2(0.001), SGD+momentum(0.9), LR=0.01, Epochs=100, Batch=64, Loss: binary_crossentropy
ANN Model B	6 layers × 16 neurons, ReLU, Sigmoid, L1L2(0.0), SGD+momentum(0.9), LR=0.01, Epochs=50, Batch=64, Loss: binary_crossentropy
ANN T (Tuned)	Layers: [128, 64, 64, 16, 128, 64], ReLU + Sigmoid, Adam, LR=0.0062, Epochs=30, Batch=64
Bayesian Network	Structure: HillClimbSearch, Scoring: K2, Inference: Variable Elimination
Random Forest	class_weight='balanced', random_state=42
XGBoost	objective='binary:logistic', eval_metric='logloss', tree_method='hist', random_state=42
LightGBM	objective='binary', class_weight='balanced', n_estimators=100, random_state=42

Table I: Models and their key hyperparameters used in evaluation and analysis.

Model	Acc.	Prec.	Rec.	F1	Time (s)
LogReg (L2)	0.91	0.91	0.91	0.91	<0.003
LogReg (L1)	0.91	0.91	0.91	0.91	<0.003
BN	0.91	0.91	0.91	0.91	7.754
ANN A	0.9107	0.9213	0.9111	0.9027	9.714
ANN B	0.9098	0.9031	0.918	0.9105	9.524
ANN Tuned	0.9087	0.9261	0.9104	0.9073	9.298
RF	0.898	0.8977	0.8982	0.8981	3.661
XGB	0.9047	0.9054	0.9047	0.9047	1.634
LGBM	0.9067	0.9072	0.9067	0.9067	0.630

Table II: Evaluation metrics and testing time for each model
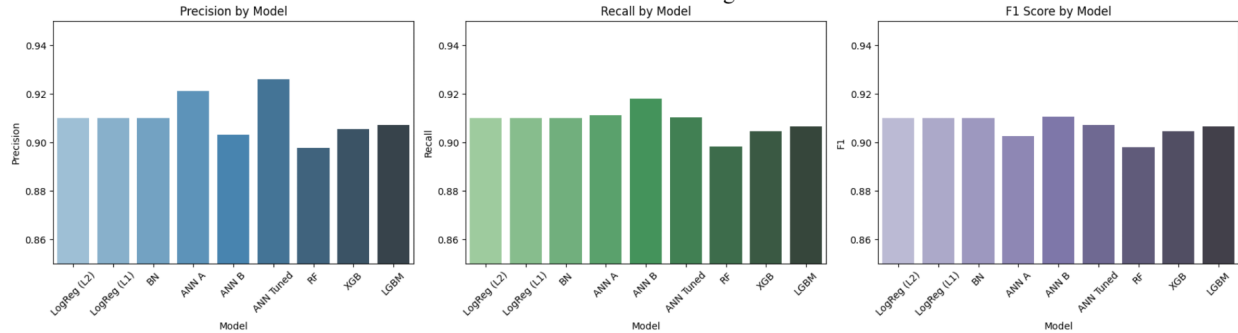
Fig 3.Performance comparison of Precision, Recall, F1 Score

	LR	ANN A	ANN B	ANN T	BN	Trees
Highest F1 / Overall Accuracy	YES	NO	YES	YES	NO	YES
Minimize False Negatives (FN)	NO (FN 740)	YES (FN 627)	NO (FN 696)	NO	NO (FN 693)	NO
Fastest Inference Time	YES (<3ms)	NO (9s)	NO (9s)	NO (9s)	NO (7.75s)	YES (0.63s)
Interpretability / Explainability	YES	NO	NO	NO	YES	Partial
Simplicity / Low Model Complexity	YES	NO	NO	NO	YES	NO
Robust to noise / non-linearity	Partial	YES	YES	YES	YES	YES
Works well with fewer features	NO	NO	NO	NO	YES	YES
Scalable for large datasets	YES	NO	NO	NO	NO	YES
Best for medical FN-critical app	NO	YES	NO	NO	NO	NO

Table III: Qualitative model evaluation

Analysis and Comparison:

The comparison table and accompanying bar plots provide a structured evaluation of several machine learning models—Logistic Regression, various Artificial Neural Networks (ANN A, ANN B, ANN T), Bayesian Network (BN), and Tree-Based models (Random Forest, XGBoost, LightGBM)—across key performance goals. These goals include maximizing F1 score and overall accuracy, minimizing false negatives (FN), achieving fast inference time, maintaining interpretability, ensuring model simplicity, handling noise and non-linearity, performing well with fewer features, scalability to large datasets, and suitability for FN-critical medical applications.

The table highlights which models fully meet each goal (YES), do not meet it (NO), or partially meet it. This visual comparison helps clarify the inherent trade-offs between model families. For instance, ANN A is the best choice for minimizing false negatives—a highly desirable property in clinical applications—while Logistic Regression offers the fastest and most interpretable solution, and Tree-Based models provide strong scalability and robustness to noise and feature redundancy. The combined bar plots of Precision, Recall, and F1 further reinforce these trends by showing clear strengths and limitations across models. Taken together, these visual tools serve as a practical guide for selecting the most appropriate model depending on the specific priorities and constraints of the deployment environment.

While Logistic Regression (L2) achieved the highest overall scores across all key quantitative metrics (Accuracy, Precision, Recall, and F1 Score: 0.9100), it did not achieve the lowest false negative rate. ANN Model A recorded the lowest false negative count (627), which is a crucial factor in medical applications where failing to identify positive cases can lead to delayed interventions or worsened outcomes.

Despite this limitation, Logistic Regression (L2) remains the most suitable model for initial deployment due to its:

- Consistently top-tier performance across all metrics
- Real-time inference speed (sub-millisecond)
- Simplicity, interpretability, and auditability for clinical review
- High recall and overall robustness across diverse patient profiles

Given the medical context, where both recall and model transparency are critical for clinical trust and adoption, Logistic Regression (L2) provides a reliable and practical baseline solution. Nonetheless, ANN Model A's superior ability to reduce false negatives suggests that it could be highly valuable as part of a two-stage screening pipeline—where a simple logistic regression first flags potential cases, and the ANN then provides more refined risk stratification. Future work could explore such hybrid approaches to maximize both accuracy and clinical utility.

VI. CONCLUSION AND DISCUSSION

This research project explored how machine learning can be leveraged for diabetes prediction using 21 health-related features from the CDC's Diabetes Health Indicators dataset. Using fundamental data cleaning, EDA, and feature selection methods, we were able to make the dataset best suited for our project. We approached the problem as a binary classification task, addressing key challenges like class imbalance, non-linear feature relationships, and feature skewness. After tuning a six-layer neural network, our model consistently achieved around 91 percent accuracy, with balanced precision and recall across both diabetic and non-diabetic classes. Our motivation for selecting this topic stems from the increasing global burden of diabetes and the lack of early, accessible diagnostic tools, especially in under-resourced regions. The model provides a foundation for automated screening tools that could be deployed in mobile apps or integrated into wearable devices, making preventive healthcare more proactive and personalized.

This project exemplifies how AI can serve as a catalyst for innovation in the healthcare sector. By analyzing behavioral, demographic, and physiological data, the model demonstrates how early warning systems can be built to reduce the risk of undiagnosed or misdiagnosed diabetes. In rural or underserved communities, such technology can help bridge the gap in medical access, allowing for earlier interventions and better health outcomes.

Deploying this model in real-world applications entails navigating challenges such as: (1) Data privacy and security to comply with regulations like HIPAA. (2) Fairness and bias mitigation to ensure the model works equitably across different population groups. (3) Clinical validation to demonstrate its reliability in real-world healthcare settings. (4) Transparency and trust. People require transparency from the machine to be able to trust it for important issues such as healthcare.

Despite these obstacles, the potential impact is significant. A well-validated, privacy-conscious diabetes risk application could aid both patients and healthcare providers, reducing the clinical burden and improving outcomes. Looking ahead, several areas present opportunities for further research: (1) Enriching features by incorporating genetic, dietary, or longitudinal health

data. (2) Improving interpretability to gain insights into feature importance and build clinician trust. (3) Developing personalized risk scores rather than binary predictions. (4) Partnering with healthcare institutions to test and refine the model in operational environments.

Our work demonstrates how a thoughtfully designed ML system can support the early detection of diabetes, potentially transforming how individuals and providers approach chronic disease prevention. By aligning machine learning capabilities with healthcare needs, this project lays the groundwork for more intelligent, accessible, and responsive medical technologies.

VII. GitHub Link, YouTube Demo Link, and Project Roadmap

The following links are clickable:

GitHub

YouTube Demo

Project Roadmap:
Starting this project, our overarching goals included training different supervised models using the CDC Diabetes Dataset to classify whether or not an individual has diabetes given various factors and deploying our best-performing model within an HTML page for the final demo. To achieve this, we created a weekly timeline and milestone roadmap to keep track of our collaborative and individual efforts, which was routinely updated during our bi-weekly meetings (online & in person). The following section details our project schedule.

Week 1: Getting Started
First, we created our GitHub repository and analyzed the data to determine any preprocessing necessities. During the exploratory data analysis (EDA), we examined the overall data distribution, outliers per feature, and evaluated dataset metrics (correlation, covariance, etc.). Based on these results, we decided on relevant features that contributed to diabetes prediction and identified a model that best fits the dataset. We were also working with an unequally weighted 3-class dataset at this stage.

Week 2: Feature Selection & Model Construction
Based on the EDA results from Week 1, we hypothesized appropriate model types for the classification task based on what we've learned in class. This included Logistic Regression, Bayesian Networks, and Artificial Neural Networks (ANN's). We started conducting literature reviews of selected health and medical research papers to understand clinical risk factors associated with diabetes and all features, regardless of feature selection results.

After our Friday meeting, we decided to split up the work by assigning each member to a specific model.

Week 3 to 6: Model Evaluation & Refinement
During this time period, each member ran and fine-tuned their models. After the first run-through (week 3), we realized that our unbalanced dataset led to biased results, where every model was classifying all data points as 'no diabetes' and having a difficult time generalizing. Because of this, we re-evaluated our dataset choice, and decided to use the alternative, equally weighted binary classification diabetes dataset (5050_split). The following week (4), model evaluations illustrated significant improvement, as predictions were now based on important patterns within the data instead of class frequency. Other model types were experimented with as well, including Bayesian, Random Forest, and XGBoost classifiers. While conducting EDA of the binary dataset (week 5), we experimented with converting all features to categorical, as only 3 predictors were continuous in the binary classifier dataset. After running evaluations on the newly created fully-categorical dataset for 21 and 13 features, we found that accuracy increased from 75% (on the non-categorical dataset) to 90% (on the fully-categorical dataset). Achieving these notable results, we moved to writing the final report and creating our demo.

Week 6-7: Finishing Details
Week 6 included writing the report, building the website, and creating our demo video. Each group member was responsible for writing their 4-5 feature literature review and reporting model methodologies/results. Additionally, two group members completed the website integration, while others wrote the abstract, introduction, conclusion, roadmap, etc. Finally, all group members collaborated to edit/polish the report, create the demo slides, and meet online for the demo recording.

References

[1] G. Roglic, "Who global report on diabetes: A summary," *International Journal of Noncommunicable Diseases*, vol. 1, no. 1, pp. 3–8, 2016. [Online]. Available: https://www.who.int/publications/i/item/9789241565257

[2] N. Sarwar, P. Gao, S. R. K. Seshasai, R. Gobin, S. Kaptoge, E. Di Angelantonio, E. Ingelsson, D. A. Lawlor, E. Selvin, M. Stampfer, C. D. A. Stehouwer, S. Lewington, L. Pennells, S. G. Thompson, N. Sattar, I. R. White, K. K. Ray, and J. Danesh, "Diabetes mellitus, fasting glucose, and risk of cause-specific death," *New England Journal of Medicine*, vol. 364, no. 9, pp. 829–841, 2010. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1008862

[3] C. Bommer, E. Heesemann, V. Sagalova, J. Manne-Goehler, R. Atun, T. Bärnighausen, and S. Vollmer, "The global economic burden of diabetes in adults aged 20–79 years: A cost-of-illness study," *The Lancet Diabetes & Endocrinology*, vol. 5, no. 6, pp. 423–430, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28456416

[4] C. M. Lambert *et al.*, *Hypertension in Diabetes*. NCBI, 2021, nCBI Bookshelf, EndoText, accessed 3.8 years ago.

[5] K. Yashi and S. F. Daley, *Obesity and Type 2 Diabetes*. Stat-Pearls Publishing, 2023, statPearls, 19 June 2023.

[6] M. Li *et al.*, "Fruit and vegetable intake and risk of type 2 diabetes mellitus: Meta-analysis of prospective cohort studies," *BMJ Open*, vol. 4, no. 11, Nov 2014.

[7] Q. Sun *et al.*, "Fruit and vegetable intake and risk of type 2 diabetes in women," *Diabetes Care*, vol. 36, no. 5, pp. 1070–1076, 2013.

[8] F. Song *et al.*, "Consumption of fruit, vegetables, and fruit juices and risk of diabetes," *Diabetes Care*, vol. 31, no. 7, 2008.

[9] P. T. Williams, "Vigorous exercise and the risk of hypertension, hypercholesterolemia, and diabetes," *Medicine & Science in Sports & Exercise*, vol. 40, no. 6, pp. 998–1006, 2008.

[10] A. L. Kennedy, R. N. Fontaine, C. R. Shuman, and E. D. Saggerson, "Relation of high-density lipoprotein cholesterol concentration to type of diabetes and its control," *British Medical Journal*, vol. 2, no. 6146, pp. 1191–1194, 1978, accessed: 2025-05-26. [Online]. Available: https://www.bmj.com/content/2/6146/1191

[11] I. Gornik, O. Gornik, and V. Gasparović, "A prospective observational study of the relationship of critical illness associated hyperglycaemia in medical icu patients and subsequent development of type 2 diabetes," *Critical Care*, vol. 14, no. 4, p. R130, 2010, accessed: 2025-05-26. [Online]. Available: https://doi.org/10.1186/cc9101

[12] X. Zhu *et al.*, "Health insurance coverage and access to care among adults with diabetes in the u.s." *American Journal of Public Health*, vol. 102, no. 12, pp. 2340–2346, 2012.

[13] R. E. Patzer *et al.*, "Health care access and affordability among adults with diabetes in the united states," *JAMA Network Open*, vol. 4, no. 12, 2021.

[14] Z. Lysy, G. L. Booth, B. R. Shah, P. C. Austin, J. Luo, and L. L. Lipscombe, "The impact of income on the incidence of diabetes: A population-based study," *Diabetes Research and Clinical Practice*, 2013, accessed: 2025-05-21. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23305902/

[15] J. Y. Choi, Y. Wang, S. Nam, and K. H. Lee, "Diabetes disparities in the united states: Trends by educational attainment from 2001 to 2020," *American Journal of Preventive Medicine*, vol. 67, no. 3, pp. 319–327, 2024.

[16] P. A. Moore, J. C. Zgibor, and A. P. Dasanayake, "Diabetes: A growing epidemic of all ages," *The Journal of the American Dental Association*, vol. 134, no. suppl. 1, pp. 11S–15S, 2003.

[17] E. A. M. Gale, "The rise of childhood type 1 diabetes in the 20th century," *Diabetes*, vol. 50, no. 12, pp. 2731–2740, 2001.

[18] J. S. Brach, J. B. Talkowski, E. S. Strotmeyer, and A. B. Newman, "Diabetes mellitus and gait dysfunction: Possible explanatory factors," *Physical Therapy*, vol. 88, no. 11, pp. 1365–1374, 2008.