Predictive Modeling for Cardiovascular Disease - CSC 240

Dominic Musumeci
dmusumec@u.rochester.edu

Yunsung Hong
yhong25@u.rochester.edu

Claire Kim
ykim144@u.rochester.edu

Joon Hyup Lee
jlee351@u.rochester.edu

***Abstract*— Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, underscoring the critical need for early detection and risk prediction. This study leverages advanced data science and machine learning methodologies to analyze a dataset comprising 70,000 patient records, aiming to predict the presence of CVD with high accuracy. Our approach integrates robust feature engineering, including the derivation of Body Mass Index (BMI) and Mean Arterial Pressure (MAP), dimensionality reduction using Principal Component Analysis (PCA), and comprehensive data standardization.**

First, we combined two Gaussian mixture models to create a binary classifier. Then, we evaluated multiple classification models, including Logistic Regression, Gradient Boosting (XGBoost and LightGBM), and Support Vector Machines (SVM). An Ensemble Learning framework combining K-Nearest Neighbors (KNN), XGBoost, and LightGBM within a logistic regression meta-model achieved the highest test accuracy of 72.80%.

To further enhance predictive performance, we explored deep learning techniques. A Feed Forward Neural Network (FFNN) with advanced optimization techniques, such as AdamW and OneCycleLR, was implemented, also achieving 72.80% accuracy. Additionally, a the data was transformed into textual input, which was then used to fine-tune BERT-base-uncased. This model acheived 70.93% accuracy.

Our findings demonstrate the effectiveness of combining data preprocessing, feature engineering, and machine learning in predicting cardiovascular disease. This work contributes to the growing field of machine learning in healthcare, offering a scalable and interpretable framework for early disease detection and prevention strategies. Future efforts will focus on incorporating additional data sources and refining model architectures to improve predictive accuracy and clinical applicability.

TABLE I

SUMMARY STATISTICS OF THE CARDIOVASCULAR DISEASE DATASET

Feature	Mean	Std Dev	Min	Max
Age (years)	53.3	6.7	29	65
Height (cm)	164.4	8.0	150	190
Weight (kg)	74.2	12.1	50	120
Systolic BP (mmHg)	128.8	15.3	90	200
Diastolic BP (mmHg)	82.4	11.9	60	120
Cholesterol Level	1.5	0.6	1	3
Glucose Level	1.2	0.5	1	3
Smoking (% smokers)	29.8%	-	0	1
Alcohol Intake (% drinkers)	24.6%	-	0	1
Physical Activity (% active)	68.9%	-	0	1
Cardiovascular Disease (% with disease)	50.0%	-	0	1

I. INTRODUCTION

This paper explores the prediction of cardiovascular disease risk using a dataset collected from routine medical examinations. The dataset includes demographic features (age, gender), clinical measurements (cholesterol, glucose levels, blood pressure), and lifestyle indicators (smoking status, alcohol intake, physical activity). The target variable (cardio) indicates whether an individual is at risk of cardiovascular disease (1) or not (0). The dataset is well-balanced, with nearly equal representation of both target classes, making it ideal for predictive analysis. We explore different machine learning and deep learning models and determine the predictive power of each. As a result we found that gradient boosting, feed forward neural networks (FFNNs), and transformers seemed to be the most effective methods. However, ultimately, we can see that ensemble methods, such as stacking classifiers and voting classifiers could have the potential.

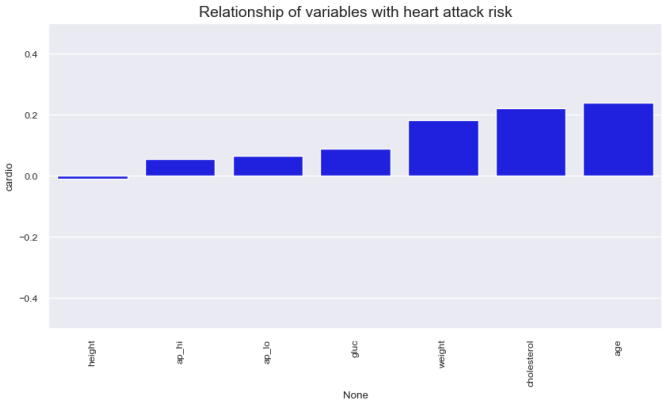

Fig. 1. Correlation between features and target variable

II. LITERATURE REVIEW

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, emphasizing the need for reliable prediction models to identify individuals at risk and enable early

intervention. Recent advancements in machine learning have introduced novel approaches that significantly outperform traditional statistical methods in predicting CVD risk. [1] This section reviews key studies that have informed the development of predictive models, highlighting their methodologies, findings, and contributions to the field.

Weng et al. [2] explored the application of machine learning models for cardiovascular risk prediction using clinical data. They compared traditional methods, such as the Framingham Risk Score, with machine learning algorithms, including neural networks and gradient-boosted models. The study demonstrated that machine learning techniques improved predictive accuracy, identifying 3.6% more high-risk patients who would benefit from preventive treatments while reducing unnecessary interventions. This research emphasized the potential of leveraging advanced algorithms to handle the complexities of CVD prediction, such as nonlinear interactions and high-dimensional data.

Similarly, Subramani et al. [3] proposed a novel stacking ensemble model that integrates base learners like Random Forest, Support Vector Machines (SVM), and Gradient Boosted Decision Trees (GBDT) with a meta-learner to refine predictions. Using SHAP values for feature selection, the study identified key predictors such as chest pain type and ST slope, achieving an impressive accuracy of 96%. This model's ability to analyze large datasets and incorporate diverse predictive features highlights its versatility and reliability in various contexts.

Both Weng et al. [2] and Subramani et al. [3] highlighted the importance of feature selection in improving model performance. Weng et al. focused on clinical variables, including age, cholesterol, and smoking status, which have been well-established in traditional models. Subramani et al., on the other hand, utilized SHAP values to rank features, identifying previously underemphasized factors like exercise-induced angina and ST slope as critical indicators. These studies underscore the necessity of tailoring feature selection techniques to the dataset and model used, ensuring the inclusion of both conventional and emerging predictors.

The integration of IoT devices and real-time data collection, as discussed in Subramani et al. [3], represents a significant advancement in the field. By leveraging data from wearable devices and electronic health records, machine learning models can capture dynamic health metrics, providing a more comprehensive risk assessment. This approach complements the findings of Weng et al., which relied on static clinical datasets, and demonstrates how emerging technologies can enhance predictive accuracy and real-world applicability.

The studies emphasize the comparative nature of predictive projects. Weng et al. showcased the superiority of neural networks and gradient boosting over traditional methods, while Subramani et al. highlighted the benefits of ensemble models, which combine the strengths of multiple algorithms. The latter's stacking approach demonstrated higher accuracy and robustness, particularly in datasets with diverse and complex features. These findings underscore the importance of evaluating multiple models to identify the best-performing approach for a given dataset.

The reviewed studies collectively illustrate the transformative potential of machine learning in cardiovascular risk prediction. By improving accuracy and uncovering complex interactions among risk factors, these models offer valuable tools for personalized medicine. Subramani et al.'s emphasis on integrating wearable device data aligns with the growing trend toward proactive health management, enabling clinicians to monitor patients continuously and intervene early. Similarly, Weng et al.'s findings highlight the feasibility of implementing machine learning in routine clinical workflows, making advanced predictive analytics accessible to a broader range of healthcare providers.

While the studies reviewed provide promising insights, they also highlight areas for improvement. Weng et al. [2] acknowledge the limitations of their dataset, which primarily consisted of static clinical data, and call for the integration of more dynamic and diverse data sources. Similarly, Subramani et al. [3] point out that their stacking model's high accuracy relies on access to comprehensive datasets, which may not always be available in real-world clinical settings. Future research should focus on expanding the availability of real-time data and developing models that can adapt to varying resource constraints. Additionally, efforts to improve model interpretability and user-friendliness will be essential for broader adoption in clinical practice.

In conclusion, the transition from traditional statistical models to machine learning-driven approaches represents a paradigm shift in cardiovascular risk prediction. Studies like those by Weng et al. [2] and Subramani et al. [3] provide a solid foundation for further research, demonstrating how advanced algorithms and innovative data sources can enhance the precision and impact of CVD prevention efforts.

III. DATA

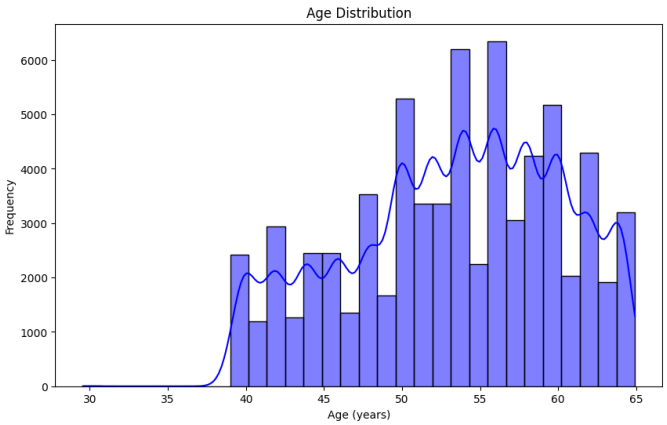

Fig. 2. Distribution of age in dataset

The Cardiovascular Disease dataset is a comprehensive collection of demographic, clinical, and lifestyle data used to predict cardiovascular disease risk. The target variable, cardio, is binary, indicating whether a patient is at risk

(1) or not (0). The dataset is well-balanced, with 50.03% of patients labeled as "No Risk" and 49.9% as "At Risk," making it ideal for building and evaluating predictive models. The age distribution reveals a middle-aged population, ranging from 29 to 64 years, with most patients concentrated between 48 and 58 years. Gender distribution shows a skewed representation, with 65.0% female and 35.0% male patients. Clinical variables such as cholesterol and glucose levels highlight potential health risks, with approximately 26% of patients having elevated cholesterol levels and 15% having above-normal glucose levels. Blood pressure data, including systolic (ap_hi) and diastolic (ap_lo) values, are critical for identifying hypertensive risks. Lifestyle factors, such as smoking (8.8%), alcohol intake (5.4%), and physical activity (80.4%), provide additional context for assessing cardiovascular risk.

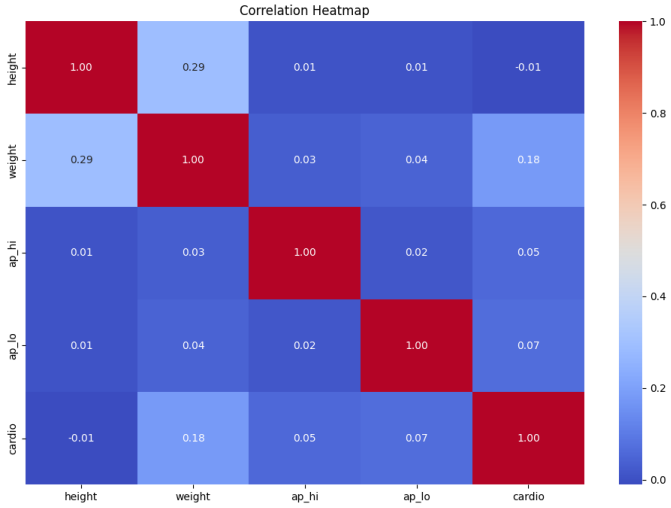

Fig. 3. Correlation heatmap of key features and cardio

Age (+0.24), cholesterol (+0.22), and weight (+0.18) show the slightly strong correlations with cardiovascular disease. The dataset's correlation heatmap highlights the relationships between variables, emphasizing the importance of both clinical and lifestyle data in understanding cardiovascular health. This dataset serves as a robust foundation for analyzing and modeling cardiovascular disease risk, offering valuable insights into the interplay of various risk factors.

IV. HYPOTHESES/GOALS

In this project, the question our study tries to answer is: what is the best model for predicting whether an individual has heart disease? We hypothesize that models incorporating lifestyle variables—such as smoking, alcohol intake, and physical activity—will enhance the ability to predict heart attack risk compared to models that use only clinical variables like age, height, and weight. We will be testing models such as Random Forest, Logistic Regression, SVM, Feed Forward Network, and Ensemble Learning. We predict that Feed Forward Networks and Ensemble Learning models will have the best outcomes because of their complexity and

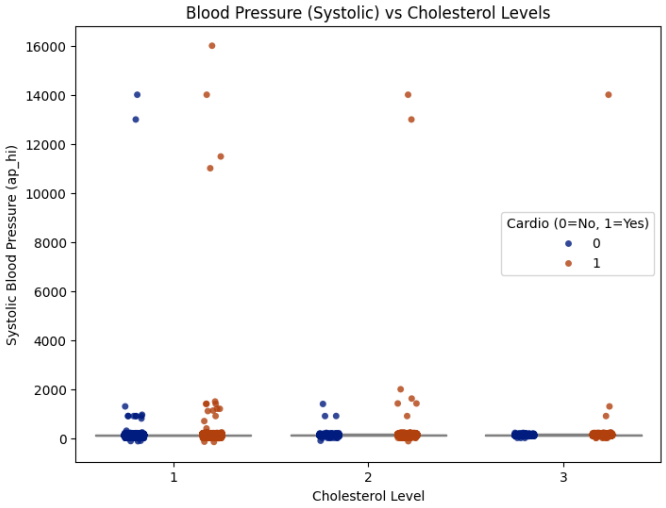

Fig. 4. Cholesterol level vs systolic blood pressure

ability to capture intricate relationships in the data. Our study aims to identify the most effective predictive model and demonstrate that a model's predictive power increases with its complexity.

V. METHODS

A. *Preprocessing*

Our analysis compared fundamental machine learning and deep learning models to predict cardiovascular disease (CVD) risk, combining feature engineering, machine learning algorithms, and statistical insights to produce meaningful results. The journey began with multiple steps of data preprocessing to remove outliers and emphasize importnace of certain attributes.

Statistical analysis was integral to understanding feature relationships. A correlation analysis revealed that age and weight exhibited moderate positive correlations with cardiovascular disease risk, while physical activity showed a weak negative correlation. These findings aligned with existing medical literature and informed the selection of features for model building.

The dataset, containing 70,000 records, required careful attention to outliers that could distort the models' predictive power. To address this, values beyond the 2.5th and 97.5th percentiles were removed. This was due to some samples included abnormally high systolic and diastolic pressures, which indicate measurement or entry errors in the dataset. Additionally, key features such as Body Mass Index (BMI) and Mean Arterial Pressure (MAP) were added to enhance the dataset's descriptive power. Both of these metrics provide standardization to attributes of one's overall health.

Categorical variables like cholesterol and glucose levels were transformed using one-hot encoding, facilitating data representation in the training and testing sets. To further transform the feature space, PCA was employed, reducing dimensionality to 15 principal components. This value was obtained through training all possible principal component

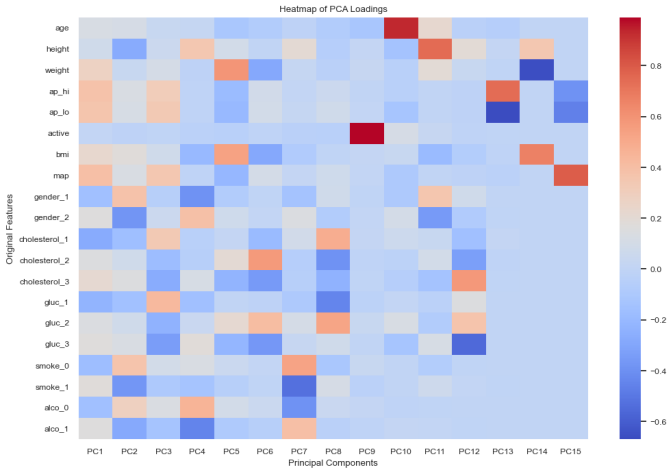

Fig. 5. Correlation of new feature space with original features

values on the dataset using fundamental models and analyzing the testing accuracy. This result is shown with the components tested on linear regression. This step not only minimized noise but also reduced multicollinearity among features, allowing for efficient and effective model training. Standardization of the data ensured that each feature contributed equally during model optimization.

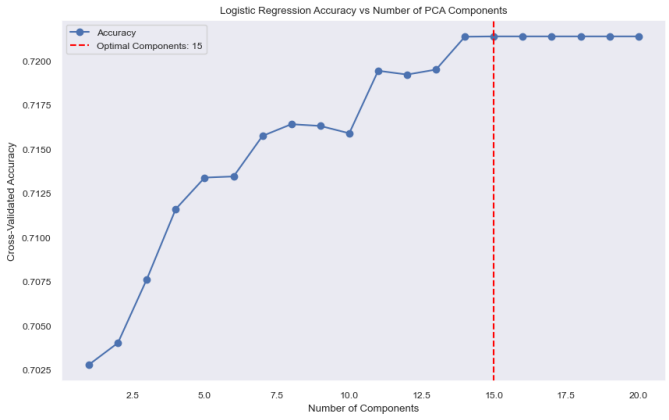

Fig. 6. Optimal number of PCA components for Logistic Regression

B. Gaussian Mixture Model

Firstly, we constructed a binary classifier using Gaussian Mixture Models. This approach demonstrates the applicability of clustering techniques in classification. We assigned class probabilities to each data point based on its likelihood under each Gaussian component. Use maximum likelihood or Bayesian inference to classify the data into the most probable class. Applying Principle Component Analysis (PCA) to the feature space with 3 components allowed this approach to obtain 69.36% accuracy.

C. Fundamental Models

For classification, we models ranging from foundational algorithms like Logistic Regression, Decision Trees, Random

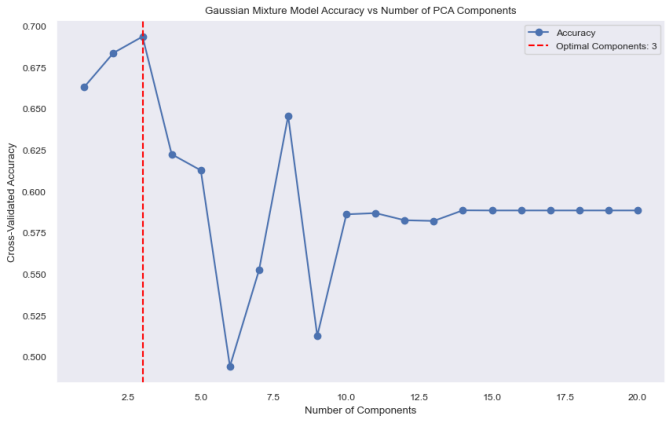

Fig. 7. Optimal number of PCA components for Gaussian Mixture Model

Forests, Support Vector Machines (SVMs), K-Nearest Neighbors (KNNs), AdaBoost, and Naive Bayes. Each model was evaluated using default hyperparameters to establish baseline performance and highlight strengths and weaknesses. Of these models, Logistic Regression seemed to demonstrate the greastest predictive power. However, one noticeable result was that the Random Forest model had the highest recall of 0.67, which is particularly important for this task. This demonstrates the model's ability to predict among positive cases, decreasing the number of false negatives and allowing for more potential cases of cardiovascular to disease to be identified.

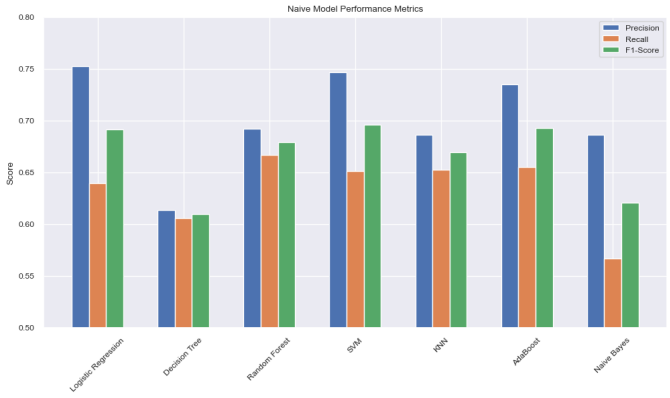

Fig. 8. Precision, recall, and f-1 scores for fundamental models

D. Gradient Boosting and Ensemble Learning

Initially, we designed a model that relied solely on gradient boosting techniques, specifically XGBoost and LightGBM, achieving 72.5% accuracy on the test set. These algorithms excel at capturing complex non-linear relationships and feature interactions through iterative learning, making them highly effective for structured data. However, to push the performance further, I implemented an ensemble learning pipeline that combined K-Nearest Neighbors (KNN) with the gradient boosting models, using a logistic regression meta-model to synthesize their predictions.

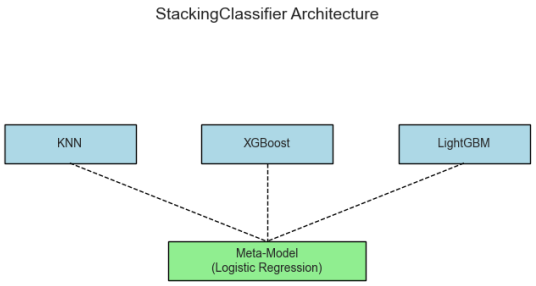

Fig. 9. Ensemble learning architecture

The addition of KNN brought the ability to capture localized patterns that gradient boosting might overlook, while the logistic regression meta-model integrated these complementary perspectives to improve the final decision boundary. Additionally, we used Grid Search to optimize the hyperparameters of the gradient boosting models. This approach increased the test accuracy to 72.8%, demonstrating the value of leveraging diverse algorithmic strengths. Additionally the meta-model provided interpretability by assigning weights to each base model's output, emphasizing their relative contributions. Although we are still only seeing slight increases in accuracy, the ensemble highlighted the importance of combining global and local learning perspectives to enhance generalization. This process underscores the effectiveness of modern ensemble techniques in building robust and lightweight predictive models.

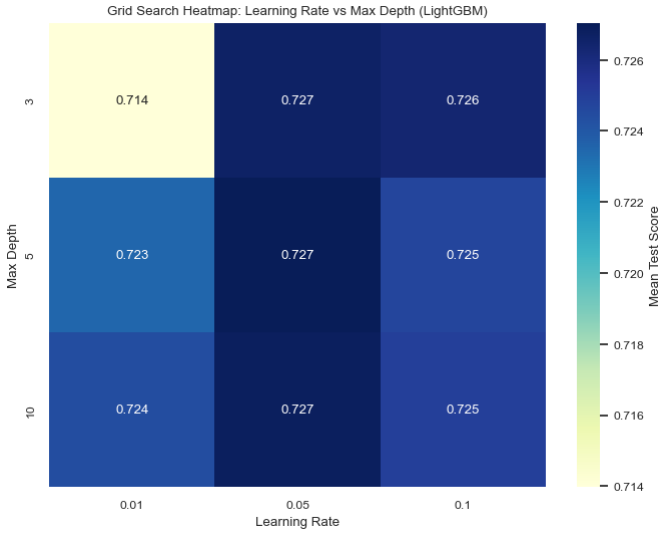

Fig. 10. GridSearch results for XGBoost

E. Feed Forward Neural Network

To push the boundaries of predictive performance, a Feed Forward Neural Network (FFNN) was implemented.

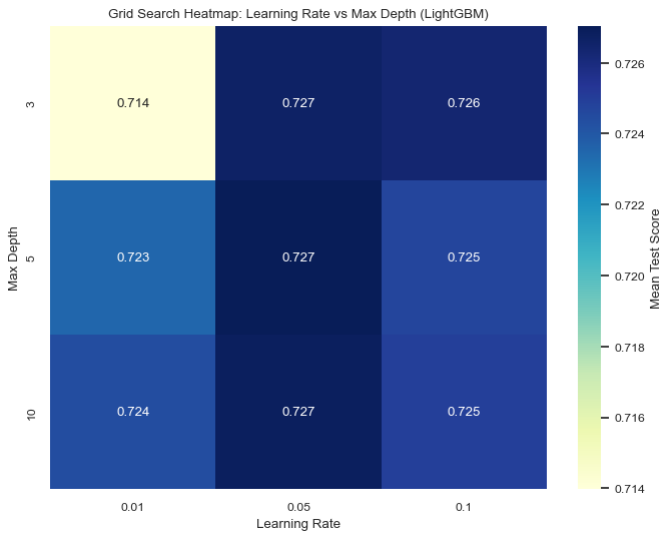

Fig. 11. GridSearch results for LightGBM

The network consisted of four fully connected layers, each utilizing LeakyReLU activations to address the vanishing gradient problem, ensuring stable gradient propagation during training. To combat overfitting, Dropout layers were strategically placed between layers, introducing regularization by randomly deactivating neurons during training. Batch normalization was applied to normalize activations across layers, stabilizing and accelerating the learning process. The

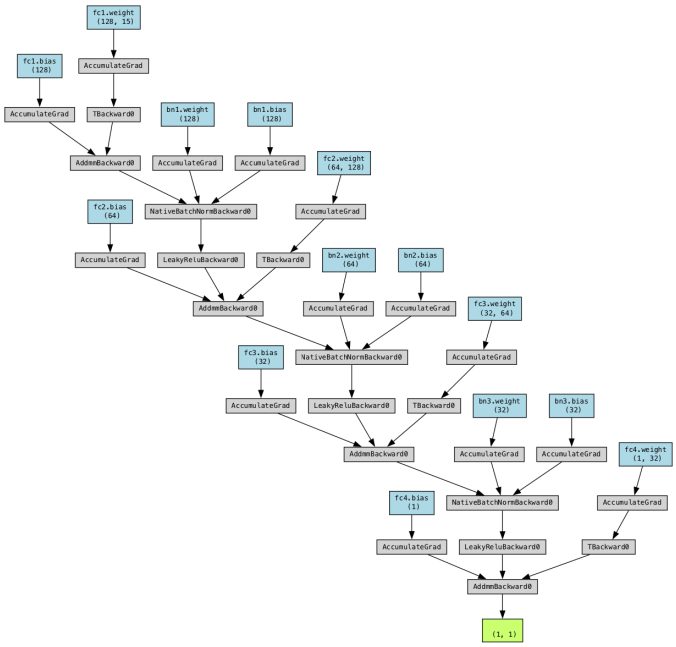

Fig. 12. Feed Forward Neural Network (FFNN) architecture

optimization strategy involved the AdamW optimizer, chosen for its effective weight decay mechanism, with a finely tuned learning rate of 0.00005 and a OneCycleLR scheduler to dynamically adjust learning rates for optimal convergence. Additionally, Xavier Uniform initialization was employed to

maintain appropriate weight scales, enhancing the efficiency of the forward and backward passes. The FFNN was trained using the binary cross-entropy loss function, ensuring robust performance in handling probabilistic outputs. Ultimately, the network achieved a test accuracy of 72.8%, effectively capturing intricate data patterns and validating its architectural and training design choices.

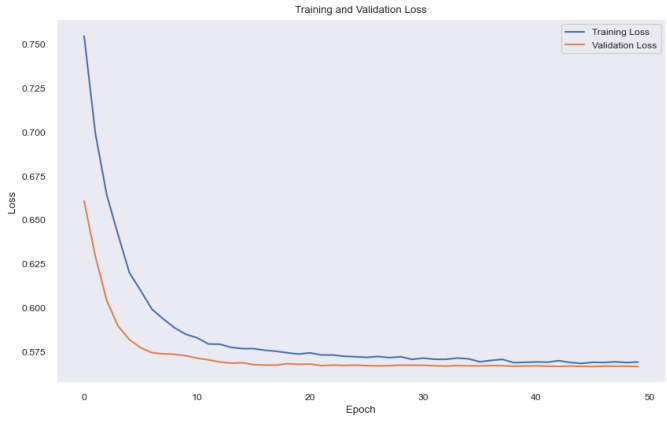

Fig. 13. Training and validation loss over training epoch

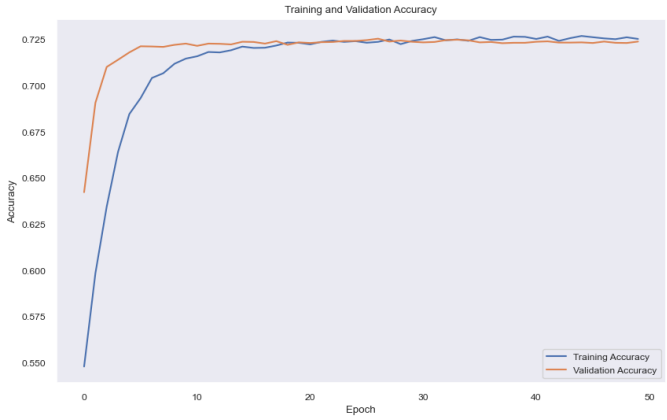

Fig. 14. Training and validation accuracy over training epoch

F. BERT Transformer

Continuing this path, we explored the application of pre-trained transformers to the task, leveraging their ability to process and understand textual data. By quantizing each feature into a label, we generated textual descriptions for each sample, transforming the dataset into a format suitable for transformers. These descriptions were used as input to BERT-base-uncased, enhanced with a binary classification head to adapt the model for our specific task. The fine-tuning process utilized training arguments that included a learning rate of 0.00002, a batch size of 16, and a weight decay of 0.01, while all other parameters adhered to default Hugging Face specifications. Despite the limited additional preprocessing applied to the dataset due to time constraints which only allowed for training on 3 epochs over the training data, the transformer demonstrated its strength in handling minimally preprocessed inputs. The fine-tuning process capitalized on BERT's contextual embeddings, capturing nuanced relationships within the transformed data. Ultimately, this approach achieved a test accuracy of 70.93%, showcasing the power of pre-trained transformers in adapting to domain-specific tasks even with constrained preprocessing efforts.

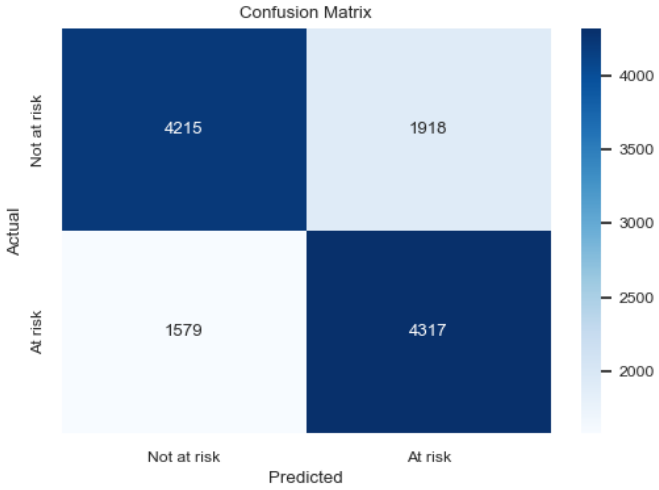

Fig. 15. Confusion matrix for Transformer model

Although this model has slightly lower overall accuracy than some of the other models, it has the highest recall. Additionally, this model was trained on less information than the others and was used to test the amount of data to make an accurate prediction. When trained on its entire textual representation, this model received an accuracy of 73.39%. There are two key results from this model. First, due to quantization applied to each category, we have created a predictive model that requires less information than the previous models. This generalizes its use cases, increasing accessibility of the model as it does not require any measurements, such as blood pressure, height, or weight, but only requires a general range for these. Second, this model had a higher recall compared to all of the other tested models. Although there was a slight decrease in accuracy compared to other advanced architectures, the transformer model had a high recall while maintaining a high enough precision. This means that given an individual with cardiovascular disease, there is a higher chance that this individual will be correctly identified using the transformer model than other model we explored. Since this correctly identifying positive cases carries more weight than negative cases, this model has the best performance in practice. Also, this indicates that there could be an effective ensemble method that combines this transformer model with a FFNN architecture, utilizing the benefits of both learning strategies.

G. Insights

The entire analysis utilized Python, leveraging libraries such as Pandas, NumPy, and SciPy for data preprocessing,

TABLE II

PERFORMANCE COMPARISON OF EACH MODEL

Metric	Logistic Regression	Decision Tree	Random Forest	SVM	KNN	AdaBoost	Naive Bayes	FFNN	Gradient Boosting	Ensemble Learning	Transformer
Accuracy	0.7209	0.6205	0.6917	0.7213	0.6843	0.7159	0.6611	0.7280	0.7252	0.7280	0.7093
Precision	0.7526	0.6134	0.6918	0.7463	0.6864	0.7349	0.6858	0.7580	0.7443	0.7438	0.6924
Recall	0.6396	0.6054	0.6666	0.6513	0.6523	0.6552	0.5663	0.6538	0.6694	0.6789	0.7322
F1 Score	0.6915	0.6094	0.6790	0.6956	0.6689	0.6928	0.6204	0.7020	0.7048	0.7099	0.7117

Scikit-learn and PyTorch for machine learning, and Matplotlib and Seaborn for visualization.

Through the integration of feature engineering, diverse machine learning methodologies, and statistical analysis, our approach not only addressed the complexities of cardiovascular disease prediction but also set a strong foundation for future improvements to our methods. Some relevant paths to consider include modifications to hyperparameters and model architectures. Possible options include combining models in a voting classifier, hyperparameter tuning to feed forward neural network, and improving text preprocessing for the transformer model. Nonetheless, this research serves as a step towards robust predictive models in healthcare. Additionally, a focus on recall could be an important consideration for this task.

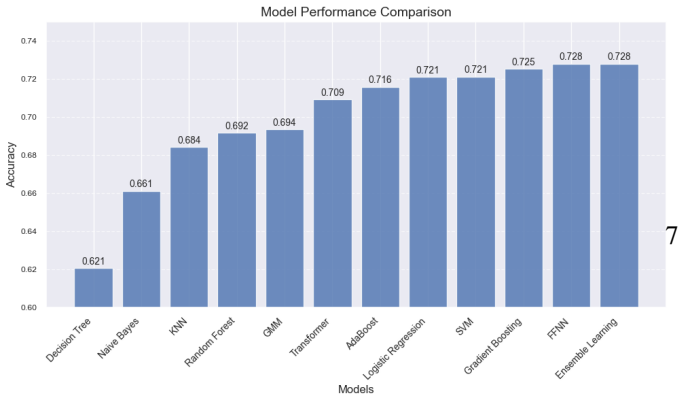

Fig. 16. Comparison of validation accuracy among all models

VI. RESULTS

Our results show that Ensemble Learning, Feed Forward Networks, and transformer outperformed other models, achieving the highest test accuracy due to their ability to capture and handle complex relationships within the data. As seen in the graph below, engineered models show increased predictive power and performance. The accuracy increases as the model becomes more complex. While our results are promising, additional preprocessing techniques could further enhance performance by refining the data and making patterns more discernible to the algorithms. Moreover, although Feed Forward Networks achieved an impressive 72.8% accuracy, focusing on hyperparameter optimization in FFNN and transformer models, test accuracy could be pushed beyond 73%.

VII. CONCLUSION

Our findings showcase the potential of advanced machine learning models in accurately predicting cardiovascular disease risk. With further refinement, these models can achieve even greater accuracy. These advancements would not only enhance the technical performance of our models but also bring us closer to impactful real-world applications.

Cardiovascular disease is the leading cause of death globally. An accurate predictive model enables early identification of high-risk individuals, allowing for timely interventions and preventive measures. Early detection can also reduce the economic burden of treating advanced stages of CVD and contribute to improving overall public health outcomes.

REFERENCES

[1] Mozaffarian, D., et al. (2015). *Heart Disease and Stroke Statistics—2015.*

[2] Benjamin, E. J., et al. (2018). *Heart Disease and Stroke Statistics—2018 Update: A Report from the American Heart Association.*

[3] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS ONE*, vol. 12, no. 4, pp. e0174944, 2017.

[4] S. Subramani, N. Varshney, M. V. Anand, M. E. Soudagar, L. A. Al-Keridis, T. K. Upadhyay, N. Alshammari, M. Saeed, K. Subramanian, K. Anbarasu, and K. Rohini, "Cardiovascular diseases prediction by machine learning incorporation with deep learning," *Frontiers in Medicine*, vol. 10, pp. 1150933, 2023.

[5] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, pp. 104672, 2021.

[6] S. Sulianova, (2018). *Cardiovascular Disease Dataset.* [Online]. Available: `https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data`