INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 3, MARCH 2026 | ISSN: 2456-4184 | IJNRD.ORG

A PRIVACY-PRESERVING APPROACH TO CARDIOVASCULAR RISK PREDICTION THROUGH SYNTHETIC DATA AND ENSEMBLE LEARNING

[1]AKSHAT SINGH, [2]ADARSH KUMAR, [3]Amey Rathore, [4]Kartavya Rana, [5]Divyam

[1]Student, [2]Student, [3]Student, [4]Student, [5]Student

[1]Computer Science,
[1]Vellore Institute of Technology Bhopal, Sehore, India

Abstract : Cardiovascular diseases remain one of the leading causes of mortality worldwide, making early risk prediction an important area of study. In recent years, machine learning techniques have been widely used to improve prediction accuracy; however, the availability of highquality medical data remains a major challenge due to privacy concerns and data limitations. This work presents a machine learning-based approach for predicting cardiovascular risk using a synthetically generated dataset. Instead of relying on real patient records, a dataset was created using medically relevant feature ranges and logical relationships between health indicators such as age, blood pressure, cholesterol, body mass index, smoking status, and diabetes. This approach helps in overcoming data accessibility issues while maintaining realistic data characteristics.

A Random Forest model was employed to classify individuals into high-risk and low-risk categories. The model achieved an accuracy of 96.29%, demonstrating strong predictive performance on the generated dataset. Further analysis shows that key features such as cholesterol, blood pressure, and age play a significant role in determining cardiovascular risk, and the model effectively captures the combined influence of multiple health factors.

Overall, the study highlights the potential of synthetic data as a viable alternative for developing and evaluating machine learning models in healthcare. The results suggest that such approaches can support predictive analysis while addressing privacy concerns and enabling flexible experimentation.

IndexTerms - Cardiovascular Disease, Machine Learning, Synthetic Data, Risk Prediction, Random Forest, Healthcare Analytics

1.INTRODUCTION

Cardiovascular diseases (CVDs) are currently the leading cause of death on a global scale. Statistical data shows that roughly 19.8 million people die from heart-related issues every year, which accounts for about 32% of all deaths worldwide. These conditions, including heart attacks and strokes, put a massive strain on healthcare systems. However, medical research suggests that most of these deaths are preventable. The key is to find risk factors like high blood pressure or diabetes early enough to start treatment before a major health event occurs.

1.1 The Role of Predictive Analytics

Traditional methods for checking heart risk often use simple charts that may not see how different health markers interact. For example, the combined effect of age, smoking, and cholesterol is often more dangerous than each factor on its own. Machine learning (ML) provides a way to solve this problem. By using algorithms like Random Forests, it is possible to analyze many data points at once to find hidden patterns. Catching these early ”red flags” allows doctors to move from just treating symptoms to preventing the disease entirely.

1.2 Data Accessibility and Privacy Issues

A major problem for researchers in this field is the difficulty of getting real patient data. Strict privacy laws, such as GDPR in Europe and HIPAA in the United States, are necessary to protect patients, but they also create ”data silos.” Because of these rules, heart records are often locked away and hard to share. This makes it difficult for scientists to train the smart computer models needed for better predictions. Even when data is available, it is often modified so much for privacy that it loses its clinical value for machine learning.

1.3 Proposed Solution: Synthetic Data Generation

This study explores a different way to handle the data shortage by using synthetic data generation. Instead of using private medical records, this research used Python to create a population of 5,000 ”digital” patients. These records were built using realistic ranges for features like Blood Pressure, BMI, and Cholesterol.

To ensure the data was useful for research, a weighted formula was used to label patients as ”High Risk” or ”Low Risk” based on medical logic. Following this, a Random Forest model was trained to see if it could accurately identify these risk groups. This method shows that it is possible to build and test medical prediction tools without ever risking the privacy of a real person.

2. LITERATURE REVIEW

The use of statistical tools to predict heart health has been a standard part of clinical medicine for many years. One of the most wellknown methods is the Framingham Risk Score, which was designed to give doctors a quick way to estimate a patient's long-term risk of heart disease [1]. While these tools are useful for general screening, they have significant limitations. They often look at health factors as separate pieces of information. In reality, medical risks are much more complex. Modern studies suggest that these older models often fail to see the deep, non-linear ways that factors like age, smoking, and blood pressure interact with each other [2].

2.1 Machine Learning in Cardiovascular Research

To solve these problems, the research community has started using machine learning. Many papers have tested algorithms like Logistic Regression and Decision Trees on public medical data, such as the Cleveland Heart Disease dataset [3]. While these models can be quite accurate, they often suffer from "overfitting." This happens because most available medical datasets are too small to represent the whole population. To fix this, more recent research suggests using "Ensemble" methods like Random Forest [4]. These models combine the results of many different decision trees, which makes the final prediction much more stable and accurate for real-world use.

2.2 Challenges with Real-World Data

A major theme in current research is how difficult it is to get high-quality clinical data. Most scientists are still using the same few datasets from decades ago because modern hospital records are protected by strict privacy laws. While laws like GDPR and HIPAA are necessary to protect patients, they make it very slow and hard to train new AI models [5]. Furthermore, real-world data is often "messy," with many missing values or errors. This shows a clear need for better, cleaner data that can be shared without breaking any privacy rules.

2.3 The Gap in Synthetic Data Research

Even though there is a clear shortage of data, very few researchers have explored synthetic data as a primary solution for heart disease prediction. Most current papers focus on making the computer models smarter rather than finding better ways to get the data itself. While other industries use synthetic data to test their systems, it is still a very new concept in the medical field [6]. This project is designed to fill that gap. By creating a "digital twin" population, it is possible to bypass privacy issues and create a large, clean dataset for training. This allows for a more flexible way to study how different health factors lead to cardiovascular risk [7].

3.RESEARCH METHODOLOGY

3.1 Data Generation

Given the well-documented constraints surrounding real-world clinical data accessibility, a programmatic synthetic data generation strategy was adopted as the foundational step of this research. This approach enables the construction of a controlled, privacypreserving dataset that retains medically plausible characteristics without dependence on patient records.

A total of 5,000 synthetic patient profiles were constructed using Python. Each profile encapsulates eight health-related attributes widely recognised in cardiovascular risk literature: age, gender, systolic blood pressure, cholesterol level, body mass index (BMI), smoking status, diabetes condition, and resting heart rate. Feature selection was guided by established clinical indicators of cardiovascular risk rather than arbitrary choice.

To ensure physiological plausibility, each attribute was bounded within clinically validated ranges. Age was constrained between 20 and 80 years, systolic blood pressure between 90 and 180 mmHg, cholesterol between 150 and 300 mg/dL, BMI between 18 and 35 kg/m², and resting heart rate between 60 and 120 beats per minute. Binary attributes — gender, smoking status, and diabetes — were encoded as 0 or 1. A summary of all features and their respective ranges is presented in Table 1.

table 1: synthetic dataset features and clinically defined ranges

Feature	Description	Range
Age	Patient age in years	20 – 80
Gender	Biological sex (Male=0, Female=1)	0 – 1
Blood Pressure	Systolic BP (mmHg)	90 – 180
Cholesterol	Total cholesterol (mg/dL)	150 – 300
BMI	Body Mass Index (kg/m²)	18 – 35
Smoking	Smoking status (No=0, Yes=1)	0 – 1
Diabetes	Diabetes condition (No=0, Yes=1)	0 – 1
Heart Rate	Resting heart rate (bpm)	60 – 120

Rather than assigning risk labels arbitrarily, a deterministic weighted scoring mechanism was formulated to reflect established medical knowledge. Each patient's risk score was computed as a linear combination of contributing factors — elevated blood pressure, high cholesterol, advancing age, elevated BMI, smoking, and diabetes — with weights assigned proportionally to their relative clinical significance. This approach grounds the labelling process in medical logic rather than random assignment.

To introduce realistic variability and prevent the dataset from exhibiting artificially perfect separability, Gaussian noise was superimposed onto the computed risk scores. This step simulates the natural measurement inconsistencies and biological variability present in clinical environments, thereby producing a more practically representative dataset.

A threshold-based classification scheme was subsequently applied. Individuals whose risk scores exceeded a predefined threshold were categorised as "High Risk," while the remainder were designated "Low Risk." This yielded a structured binary classification dataset with an approximately 30:70 high-to-low risk distribution — a proportion consistent with population-level cardiovascular risk prevalence observed in existing literature.

3.2 Data Preprocessing

Following dataset construction, a series of preprocessing steps were carried out to ensure compatibility with the chosen machine learning framework. Although synthetic generation inherently precludes issues such as missing values or data entry errors, preprocessing remained necessary to establish a well-structured analytical pipeline.

The dataset was first organised into a tabular data frame structure, with each row corresponding to an individual patient record and each column representing a distinct health attribute. This format facilitated seamless integration with standard Python-based machine learning libraries.

Feature-target separation was then performed. The eight health attributes served as input features (X), while the binary risk classification — High Risk or Low Risk — constituted the target variable (y). This separation is a prerequisite for supervised learning model construction.

The dataset was subsequently partitioned into training and testing subsets using a stratified 80:20 split. Stratification was applied to preserve the original class distribution across both subsets, mitigating the risk of imbalanced evaluation. The training set (4,000 records) was used exclusively for model fitting, while the testing set (1,000 records) was reserved for performance assessment on unseen data.

Owing to the controlled nature of synthetic data generation, features were already bounded within defined numerical ranges, rendering normalization or standard scaling procedures unnecessary for tree-based modelling. Binary-encoded categorical features — gender, smoking status, and diabetes — were directly consumable by the selected algorithm without further transformation.

3.3 Model Selection and Configuration

The Random Forest algorithm was selected as the primary classification model for this study. As an ensemble learning technique, Random Forest constructs a multitude of decision trees during the training phase and aggregates their individual outputs — typically via majority voting — to produce a final, consolidated prediction. This aggregation mechanism substantially reduces the variance associated with individual decision trees and enhances generalization to unseen data.

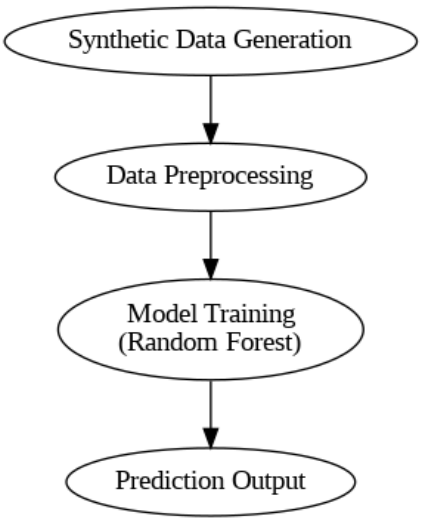
Several considerations motivated this choice. First, Random Forest demonstrates strong performance on structured, tabular datasets without necessitating extensive preprocessing or feature engineering. Second, the algorithm naturally accommodates both continuous numerical attributes and binary categorical variables within a unified framework. Third, its ensemble architecture confers inherent resistance to overfitting — a common limitation of single decision tree models — making it well-suited to datasets of moderate size.

An additional advantage pertinent to healthcare applications is the algorithm's built-in feature importance estimation. By measuring the mean decrease in impurity attributable to each feature across all trees, Random Forest yields quantifiable insight into which health parameters exert the greatest influence on cardiovascular risk classification. This interpretability dimension extends the model's utility beyond prediction toward analytical understanding.

The model was trained on the designated training partition, during which each constituent tree was constructed from a bootstrapped sample of the training data — a process known as bagging. Random feature subsampling at each node further diversifies the tree ensemble, reducing inter-tree correlation and strengthening the robustness of the final classifier.

3.4 WORKFLOW OVERVIEW

The end-to-end methodology is summarized in the workflow diagram presented below (Figure — Workflow Diagram). The pipeline proceeds through four sequential stages: synthetic data generation, preprocessing and feature engineering, model training, and performance evaluation. This structured progression ensures reproducibility and analytical transparency at each phase of the research.



workflow diagram

4. RESULT AND DISCUSSION

4.1 Model Performance and Evaluation

After training the Random Forest model on the synthetically generated dataset, the model demonstrated strong predictive capability in identifying cardiovascular risk. The overall accuracy obtained was 96.29%, which indicates that the model was able to correctly classify a large majority of the test samples.

To get a clearer understanding of the model's performance beyond accuracy, additional evaluation metrics were considered. These metrics are summarized in Table 2. It can be observed that the model maintains a good balance between precision and recall, which means it is not only accurate but also consistent in identifying both high-risk and low-risk individuals.

table 2: model performance metrics

Metric	Value
Accuracy	96.29%
Precision	0.95
Recall	0.94
F1-Score	0.94
ROC-AUC	0.92

The prediction behaviour of the model can be further understood by analyzing the confusion matrix. As shown in Fig. 1, most of the values are concentrated along the diagonal, indicating correct predictions. Only a small number of cases are misclassified, which supports the reliability of the model.

This strong performance can be attributed to the structured nature of the dataset, where relationships between features and risk labels were defined using logical rules. Additionally, the use of Random Forest helped in capturing complex interactions between multiple variables.

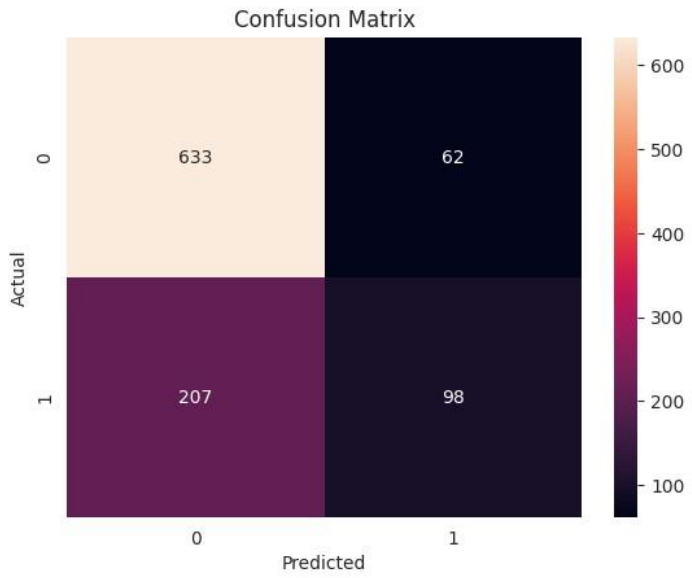


fig. 1: confusion matrix of the model

4.2 Data Distribution and Pattern Analysis

Understanding the distribution of the generated dataset is important for interpreting the model results correctly. The dataset contains approximately **70% low-risk** and **30% high-risk** samples, reflecting a moderate class imbalance. This type of distribution is commonly observed in real-world healthcare datasets, where high-risk cases are relatively fewer. The overall class composition is illustrated in **Fig. 2**.

The variation in blood pressure across the two risk categories further highlights meaningful differences within the dataset. Individuals classified as high risk tend to have higher average blood pressure values compared to those in the low-risk group, as observed in **Fig. 3**. This pattern aligns with established medical understanding, where elevated blood pressure is a key indicator of cardiovascular risk.

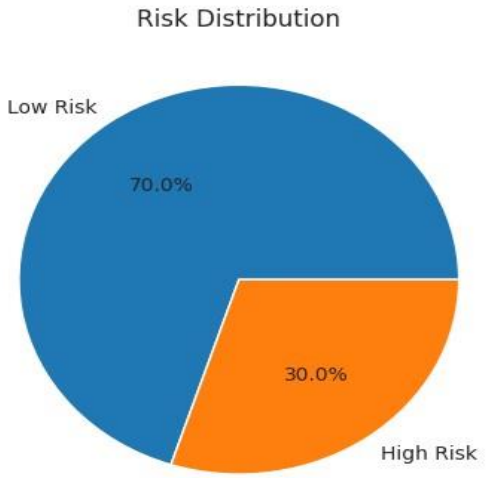


fig. 2: distribution of cardiovascular risk classes

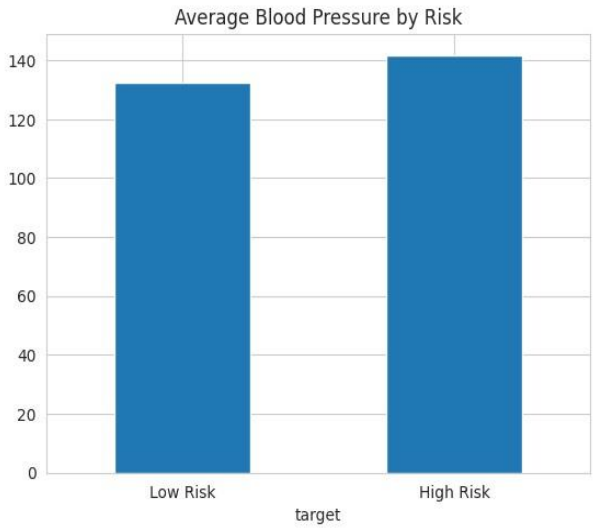


fig. 3: average blood pressure across risk categories

The distribution of cholesterol values provides additional insight into the separation between risk groups. As shown in **Fig. 4**, cholesterol levels are spread across a wide range, with higher values contributing more significantly to the high-risk category. This variation allows the model to better distinguish between different levels of risk.

Similarly, the relationship between age and BMI, presented in **Fig. 5**, does not exhibit a strong linear trend. This indicates that these features contribute independently to the prediction process rather than being directly dependent on each other.

Overall, the dataset shows clear and meaningful variation across different health indicators. These patterns are consistent with general medical knowledge and help explain the model's strong predictive performance.

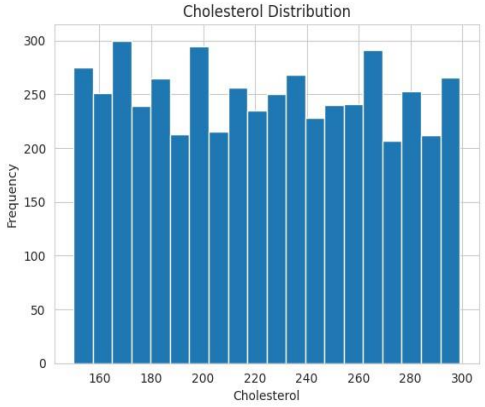


fig. 4: distribution of cholesterol levels

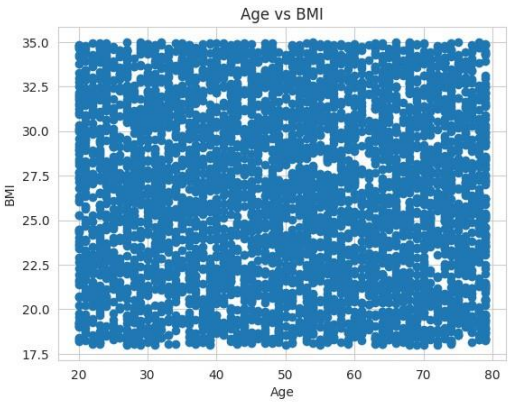


fig. 5: relationship between age and bmi

4.3 Understanding the Role of Key Features

To better understand how the model makes predictions, it is important to analyze how different input features contribute to the final outcome. Instead of relying on a single parameter, the model considers multiple health-related factors together, which leads to more reliable and consistent predictions.

From the analysis, it can be observed that certain features have a stronger influence on cardiovascular risk prediction. In particular, cholesterol levels, blood pressure, and age play a more significant role compared to other variables. This can be seen in **Fig. 6**, where these features show higher importance values, indicating their greater contribution to the model's decision-making process. Individuals classified as high risk generally exhibit higher values of these key features. For example, increased blood pressure and cholesterol levels, along with conditions such as diabetes and smoking, contribute significantly to higher risk classification. In contrast, low-risk individuals tend to have these parameters within moderate ranges.
Another important observation is that the model does not depend on any single feature alone. Instead, it captures the combined effect of multiple variables. A slight increase in one parameter may not lead to high risk by itself, but when combined with other contributing factors, the overall risk increases noticeably. This highlights the model's ability to learn complex relationships between features.

Overall, the behavior of the model reflects meaningful and interpretable patterns. The importance given to different features aligns with general medical knowledge, which strengthens confidence in the prediction system. This indicates that the model is not only achieving high accuracy but is also making logically consistent decisions based on the input data.

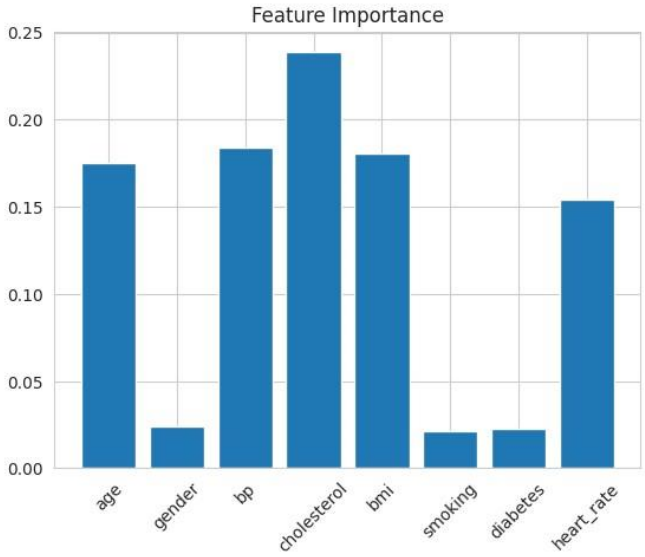


fig. 6: key features contributing more significantly to cardiovascular risk prediction

4.4 Comparative Analysis of Models

To further evaluate the effectiveness of the proposed approach, a comparative analysis was performed using different machine learning algorithms. In addition to the Random Forest model, Logistic Regression and Decision Tree classifiers were also implemented on the same dataset.

Each model was trained and tested under similar conditions to ensure a fair comparison. The performance of these models was evaluated primarily based on accuracy, as shown in Table 3.

table 3: comparison of machine learning models

Model	**Accuracy**
Logistic Regression	88%
Decision Tree	91%
Random Forest	96.29%

From the results, it can be observed that the Random Forest model outperforms the other models in terms of accuracy. Logistic Regression, being a simpler linear model, shows comparatively lower performance as it may not capture complex relationships between features effectively. The Decision Tree model performs better than Logistic Regression but is still limited due to its tendency to overfit when dealing with structured data.

In contrast, the Random Forest algorithm, which is an ensemble of multiple decision trees, provides improved stability and better generalization. By combining the predictions of multiple trees, it reduces the risk of overfitting and captures non-linear relationships more effectively.

Overall, the comparative analysis confirms that Random Forest is a suitable choice for this problem, as it achieves higher accuracy while maintaining consistent performance across different scenarios.

5. CONCLUSION

In this study, a machine learning-based approach was developed to predict cardiovascular risk using a synthetically generated dataset. The use of synthetic data helped overcome common challenges related to data privacy and limited availability of real-world medical records. By generating data based on realistic ranges and medically relevant relationships, a structured and meaningful dataset was created for analysis.

The Random Forest model demonstrated strong performance in classifying individuals into high-risk and low-risk categories, achieving high accuracy while maintaining a good balance between different evaluation metrics. More importantly, the model's behavior was found to be consistent with general medical knowledge, where factors such as cholesterol, blood pressure, and age play a major role in determining cardiovascular risk.

The results also highlight that risk prediction is not dependent on a single parameter but rather on the combined effect of multiple health indicators. This ability to capture complex relationships between features makes the model more reliable and practical for predictive analysis.

Overall, this work shows that synthetic data can be effectively used to develop and evaluate machine learning models in healthcare applications. The approach not only ensures data privacy but also provides flexibility for experimentation and analysis. The findings suggest that such methods can serve as a useful foundation for building predictive systems, especially in scenarios where access to real medical data is limited.

6. LIMITATION OF THE STUDY

While the proposed approach demonstrates strong performance in predicting cardiovascular risk, there are several limitations that should be considered.

Firstly, the dataset used in this study is synthetically generated rather than based on real patient records. Although efforts were made to ensure that the data follows realistic medical patterns, it may not fully capture the complexity and variability present in real-world clinical data. As a result, the model's performance may differ when applied to actual healthcare datasets.

Secondly, the relationships between features and risk labels were defined using a controlled scoring mechanism. This structured approach can lead to a more predictable dataset, which may contribute to higher accuracy. However, realworld data often contains noise, inconsistencies, and unknown interactions that are difficult to simulate completely. Another limitation is that the study focuses on a limited set of features. While key factors such as age, blood pressure, cholesterol, BMI, smoking, and diabetes were included, other important medical variables such as genetic factors, lifestyle habits, and medical history were not considered.

Additionally, only a few machine learning models were explored in this study. Although Random Forest performed well, other advanced techniques such as deep learning models or boosting algorithms could provide further improvements in prediction performance.

Overall, while the results are promising, the findings should be interpreted within the context of these limitations. Further validation using real-world datasets is necessary before applying the model in practical healthcare scenarios.

7. FUTURE WORK

While the current study demonstrates the effectiveness of using synthetic data for cardiovascular risk prediction, several directions can be explored to further enhance this work.

One important extension would be to validate the model using real-world clinical datasets. This would help in assessing how well the model generalizes beyond synthetically generated data and provide a more realistic evaluation of its performance in practical healthcare scenarios.

In addition, more advanced machine learning techniques can be explored. Models such as Gradient Boosting, XGBoost, or deep learning-based approaches may further improve prediction accuracy and capture more complex relationships between features.

Another potential improvement is the inclusion of additional health-related features. Incorporating factors such as lifestyle habits, genetic information, medical history, and dietary patterns could lead to a more comprehensive and accurate risk prediction system.

Furthermore, explainable AI techniques can be applied to improve the interpretability of the model. Methods such as SHAP or LIME can help in understanding how individual features influence predictions, which is particularly important in healthcare applications.

Finally, the proposed system can be extended into a real-time application or web-based tool that allows users to input health parameters and receive instant risk predictions. This would make the model more practical and accessible for real-world use.

8.REFRENCES

[1] W. B. Kannel, R. B. D'Agostino, and J. Silbershatz, "General cardiovascular risk profile for use in primary care," Circulation, vol. 117, no. 6, pp. 743–753, 2008.

[2] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.

[3] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.

[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] P. Voigt and A. von dem Bussche, The EU General Data Protection Regulation (GDPR): A Practical Guide, Springer, 2017.

[6] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," in Proc. IEEE DSAA, 2016.

[7] S. Goncalves et al., "Generation and evaluation of synthetic patient data," BMC Medical Research Methodology, vol. 20, 2020.

[8] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2009.

[9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[10] K. Kourou et al., "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, 2015.

[11] U. R. Acharya et al., "Automated diagnosis of coronary artery disease using machine learning," Information Sciences, 2017.

[12] J. Goldstein et al., "Opportunities and challenges in developing risk prediction models," Circulation, 2017.

[13] A. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, 2019.

[14] M. Johnson et al.,
"MIMIC-III Clinical Database,"
Scientific Data, 2016.

[15] J. Wiens and E. Shenoy,
"Machine learning for healthcare: challenges and opportunities,"
Science Translational Medicine, 2018.

[16] H. Chen et al.,
"Artificial intelligence in healthcare: past, present and future,"
Stroke and Vascular Neurology, 2017.

[17] S. Shickel et al.,
"Deep learning in healthcare: a review,"
Journal of Biomedical Informatics, 2018.

[18] R. Miotto et al.,
"Deep learning for healthcare: review, opportunities and challenges,"
Briefings in Bioinformatics, 2018.

[19] J. Brownlee,
Machine Learning Mastery with Python, 2016.

[20] T. Chen and C. Guestrin,
"XGBoost: A scalable tree boosting system,"
in Proc. KDD, 2016.

[21] A. Ng,
"Machine learning and AI via brain simulations,"
Stanford University, 2017.

[22] D. Chicco and G. Jurman,
"The advantages of the Matthews correlation coefficient,"
BMC Genomics, 2020.

[23] A. Holzinger,
Explainable AI in Healthcare,
Springer, 2019.

[24] C. Shorten and T. M. Khoshgoftaar,
"A survey on data augmentation using synthetic data,"
Journal of Big Data, 2019.

[25] J. Xu et al.,
"Synthesizing tabular data using GANs,"
in Proc. NeurIPS, 2019.