



OPEN Kolmogorov–Arnold Networks for predicting carotid intima-media thickness in cardiovascular risk assessment

Ali Al Bataineh¹, Bandi Vamsi², Mohammed El-Abd³✉ & Bhanu Prakash Doppala⁴

Carotid Intima-Media Thickness (CIMT) is defined as a non-invasive and well-validated sign of asymptomatic atherosclerosis and an early predictor of cardiovascular disease (CVD). We assembled a carefully curated dataset of 100 adult patients, encompassing 13 clinical, biochemical and demographic variables routinely collected in outpatient practice. After a five-stage pre-processing pipeline median/mode imputation, categorical encoding, Min–Max scaling, inter-quartile-range outlier removal and SMOTE-NC balancing we trained a Kolmogorov–Arnold Network (KAN) to assign each patient to one of four CIMT-defined risk tiers mentioned as “No”, “Low”, “Medium”, “High”. Feature-selection tests (Spearman, Pearson, ANOVA and χ^2) removed redundant predictors and improved interpretability. The KAN, implemented with ELU-activated hidden layers and a Softmax output was benchmarked against six conventional algorithms like Support Vector Machine, Decision Tree, Logistic Regression, Stochastic Gradient Descent, Deep Neural Network, Random Forest and Multi-Layer Perceptron. On stratification of five-fold cross-validation the proposed model achieved 93% accuracy, 93% precision, 93% recall, 91% F1-score and a ROC-AUC of 0.97, outperforming all baseline models by 8–19%. These results demonstrate that KAN’s capacity in capturing arbitrary connections and handling multi-class tasks demonstrating its potential as a low-cost and promising tool for early cardiovascular risk hierarchy.

Keywords Kolmogorov–Arnold network (KAN), Cardiovascular risk prediction, Carotid Intima-Media thickness (CIMT), Machine learning models

Cardiovascular health is commonly evaluated by Carotid Intima-Media Thickness (CIMT)¹. It is used to estimate the inner layer thickness present inside the walls of carotid artery by giving a non-invasive identifier of atherosclerosis and the total vascular health^{2,3}. If the CIMT value is high, it is related with the highest increase in the cardiovascular disease risk (CVDs) leading to strokes and heart attacks³. Therefore, CIMT has become a critical parameter in identifying the problem in advance and to prevention methods which are suggested for the cardiovascular risk^{4,5}.

Variations in lifestyle, genetic predisposition and hormonal changes are the main causes of the risk. Most men have a higher tendency of CIMT values when compared to that of women⁶. For instance, the world-wide data shows a mean CIMT of 0.8 mm value in men whereas 0.7 mm value among middle-aged women^{7,8}.

Most deaths worldwide are caused by cardiovascular diseases. Based on World Health Organization (WHO) 2020 report about 17.9 million deaths that happen annually are due to this cardiovascular disease^{9,10}. About 28% of the deaths in India are due to CVD, with an increasing rise of CIMT irregularities based on the different lifestyle variations, larger incidents in metabolic changes like obesity, diabetes and urbanization¹¹. With an abnormal CIMT values, the patients have a 2–3 times higher risk due to cardiovascular mortality worldwide¹². The limited healthcare measures and late detection became a problem in identifying the cardiovascular disease in India¹³.

With the rapid evolution of techniques, particularly in the field of Artificial Intelligence (AI) and Machine Learning (ML) have many advantages in evaluating the cardiovascular risk^{14,15}. These are done through enabling various techniques to analyze the complicated methods in the day-to-day life change, clinical data to identify

¹Artificial Intelligence Center, Norwich University, Northfield, VT 05663, USA. ²Department of Artificial Intelligence, Madanapalle Institute of Technology & Science, Madanapalle 517325, Andhra Pradesh, India. ³College of Engineering and Applied Sciences, American University of Kuwait, Salmiya, Kuwait. ⁴Data Analytics, Generation Australia, Sydney 2000, Australia. ✉email: melabd@auk.edu.kw

the results like thickness in CIMT and risk of cardiovascular disease¹⁶. The processing of larger datasets is done through ANN models which mimics the structure of human brain functioning and gather non-linear relations by giving accurate identifications for CIMT diagnosis and the related risks^{17,18}. In order to define this through simple components, a novel KAN method in the theory of mathematics is used to combine multiple variant features. This model excels in functioning of non-linear relation along with equalizing the interpretability by making it as a strong component for the identification of CIMT risk levels. The major aims of this study are to:

1. Developing a Kolmogorov–Arnold Network (KAN) to identify the accurate Carotid Intima-Media Thickness (CIMT), leveraging its unique ability in model for complicated non-linear relationships.
2. Analysing the KAN performance over modern ML models and ANN models to initiate its superiority.
3. Describing a novel composing methodology based on KAN identifications to divide the patients into different Risk categories like: 'no', 'low', 'medium' and 'high'.
4. Spotlight the feasibility of combining KAN-based CIMT prediction into routine cardiovascular Risk assessment practices.

The remaining sections in the work are categorized as follows, Sect. 2: Related Works. This section discusses the existing research on CIMT prediction, highlighting the limitations of traditional ML and ANN models. Section 3: Methodology, describes the dataset, pre-processing steps, and techniques of feature selection like chi-square tests, correlation analysis, ANOVA. It also describes the details on the designing and applications of the KAN model. Section 4: Results and Discussions, this showcases the performance of KAN and compare it with other models using metrics like Accuracy, MSE, R^2 , and AUC-ROC. Section 5: Conclusion and Future Scope, to discuss about the current work limitations and proposes directions for future scope of research, such as integrating larger datasets or applying KAN in medical domains.

Related works

This section portrays the upgraded developments in Artificial Intelligence (AI) and Machine Learning (ML) for predicting CIMT, an indispensable factor for cardiovascular health. It delves deeper into the application of numerous traditional and Deep Learning (DL) models by analysing their strengths, limitations and spot the cracks in current research.

Y. Shin et al.¹⁹, suggested an approach to automate the analysis of Carotid Intima-Media Thickness (CIMT) from ultrasound videos with the help of Convolutional Neural Networks (CNNs). To detect the explicate space within the designated frames where CIMT measurement is performed. A CNN is optimized in classification of frames into end-diastolic and non-end-diastolic categories established on temporal and spatial features. This work can develop enormous volumes of ultrasound data proficiently, making it advisable for large-scale cardiovascular.

Patino-Alonso et al.²⁰ organized a prevalence study associated with 501 Spanish individuals aged from 35 to 75 years all with no history of cardiovascular disease. They measured carotid-femoral pulse wave velocity with an ultrasound machine. To predict the effect of micro and macronutrients on improved vascular aging, the researchers used ensemble machine learning model which integrates multiple algorithms to optimize predictive accuracy. By scrutinizing both macro and micronutrients, the research provides a wide-range view of dietary influences on vascular health. The cross-sectional design restricts the capability to infer causality.

Ben Tekaya et al.²¹ incorporates 47 patients diagnosed with SpA, 47 age-matched and sex-matched healthy controls. The researchers employed CIMT ultrasound-based approach to measure the thickness in the walls. It functions like indicator for atherosclerosis. The research exposed that patient with SpA, even in the nonexistence of traditional cardiovascular Risk factors, exhibited signs of endothelial dysfunction and increased CIMT compared to healthy individual. These outcomes suggest a significant Risk of subclinical atherosclerosis in this patient population. Further the study is necessary for elucidating the underlying mechanisms associating SpA with endothelial dysfunction and escalated CIMT.

Sudha, S. et al.²² analyzed diverse physiological and biochemical parameters, incorporating CIMT, which functioned as a significant biomarker in their analysis. By emphasizing on the Indian population, the research addresses a demographic significant challenge of CVDs, presenting focused insights for public health interventions. Employing CIMT elevates the predictive precision for CVD onset, as it is a time-tested indicator of atherosclerosis. The study fails to specify the machine learning models used for their validation methods. Future research should consist extensive overview of the models and validation techniques to ensure reproducibility and reliability.

Johri, A. et al.²³ undertook a study engaging 459 individuals who underwent an ultrasound on focused carotid B-mode, contrast-enhanced and coronary angiography. The productivity of these ML models was evaluated against conventional statistical methods, entailing multivariate and univariate analyzes for prediction of CAD also for CV result presumption of the Cox proportional hazard model is also used. ML models, specifically RF and RSE, exhibited superior performance over conventional statistical methods in predicting CAD and CV results. The use of carotid ultrasound features furnishes a non-invasive means of assessing cardiovascular Risk. The review period was limited to 30 days. Long-term revision is vital in validating the performance of ML models over extended periods.

The core outcome was functional disability assessed by the modified Rankin scale (mRS) with a three-month of post-hospital admission. The Machine learning models along with neural networks are employed in identifying short-term disability and mortality built on the collected attributes. The findings are predicted on a specific cohort. Validating the estimative models across diverse populations would elevate their generalizability. The study²⁴ focuses on short-term outcomes (three months). Investigating the estimative value of CIMT and NIHSS for long-term outcomes would provide additional extensive knowledge.

Bhagawati, M et al.²⁵ carried out a study engaging 459 participants who underwent coronary angiography, ultrasound on contrast-enhanced and focused carotid B-mode. This research utilized eight distinct models of DL to evaluate their efficacy in anticipating CAD Risk and CV events. Both univariate and multivariate analysis were conducted to identify significant Risk predictors. DL models sometimes works as “black boxes” for making it intricate in understanding their method of decision-making. Enhancing the interpretability of these models is vital for clinical adoption.

Prabha, P. et al.²⁶ utilized a dataset comprising B-mode ultrasound images to measure CIMT, an acknowledged marker for cardiovascular diseases. They applied various ML and DL techniques to analyze the CIMT measurements and categorized the risk of infarction. The study involved pre-processing the ultrasound images to pull out relevant features, subsequent to training and testing the aforementioned models to analyze their performance in Risk classification. The ability of CNNs to involuntarily remove categorized features from raw image data contributed to their superior performance.

Lakshmi Prabha et al.²⁷ organized a study engaging 110 participants, comprising 55 T2DM subjects and 55 non-diabetic controls. CIMT measurements were acquired using ultrasound imaging. Utilizing ultrasound imaging for CIMT measurement equips a non-invasive method for assessing cardiovascular Risk. This application employs transfers learning techniques, especially the VGG-16 model, significantly elevating the accuracy of CVD Risk prediction.

While most existing studies^{19–27} concentrate on image-based methods for CIMT prediction, this work shifts the focus to structured clinical and demographic data in .CSV format. This approach addresses critical challenges such as computational complexity, data accessibility, and model interpretability. By leveraging .CSV data, we show how data is used to measure cardiovascular risk holistically using easily accessible attributes, offering a useful and effective substitute for image-based techniques. The combined use of these methods can be investigated in future studies to achieve even more robust predictions.

Image-based approaches concentrate mostly on CIMT measurement and segmentation. However, cardiovascular disease (CVD) risk is influenced by a variety of clinical, demographic and lifestyle factors that are frequently encoded in structured data formats. .CSV files enable the incorporation of variables such as age, BMI, smoking habits, and so on, providing a more thorough approach to CVD risk prediction. This research aims to evaluate how specific clinical characteristics influence CIMT changes and overall cardiovascular Risk. Such an analysis is fundamentally feature-driven; hence tabular data is the best option.

The research gap we identified, training deep learning models on images requires huge computational resources (e.g., GPUs) as well as large amounts of annotated data. In contrast, Machine learning models trained on .CSV files are computationally light weight and capable of producing reliable findings even on small datasets.

Methodology

This section explains the context of the problem, specific features and data availability. The image-based methodologies explain the importance in Carotid Intima Media Thickness (CIMT) diagnosis. With the use of structured data in CSV format that gives clear advantages that orients the features of this study. The methodology flow of this study is structured to guide the reader through a detailed way from raw data acquisition to model deployment and evaluation. The dataset description is given under Sect. 3.1, followed by pre-processing and feature engineering steps are mentioned under Sect. 3.2 to 3.3. The Sect. 3.4 combines theoretical background of Kolmogorov–Arnold Network (KAN), while Sect. 3.5 presents its architectural implementation. Optimization strategies and algorithmic workflow are explained in Sect. 3.6 and 3.7. This structure ensures that every phase of the study is presented in context and builds upon the former one for improved readability and clarity.

Dataset

The main aim of the study relies on demographic and clinical structured data. This data²⁸ is taken as main predictors also by enabling accurate and interpretable results for the diagnosis of CIMT. The dataset used in this study includes demographic, biochemical and clinical data relating to cardiovascular health which mainly aims on CIMT. The dataset used is organized to entitle effective diagnosis of different level of risks in CIMT with the help of Kolmogorov–Arnold Networks (KAN) and machine learning models. A total of 100 records, which represent diverse clinical and demographic characteristics. The dataset features used in this work are tabulated under Table 1. The dataset is classified to testing and training phases for validation of the model, nearly 80% of the dataset is reserved for cross-validation and training. The remaining 20% of it is used for the final testing purpose.

The dataset²¹ mentioned in this work is available in public and also licensed source, hosted on the ‘Mendeley Data repository’. The identifiers of the patients were removed, and the dataset is completely anonymized before implementing in the model with standard ethical guidelines for human data use. There is no personally identifiable information mentioned and the data is labelled in such a way that ensures the privacy protection.

Data pre-processing

An important phase in which the dataset is ensured and made available for machine learning modelling and evaluation is performed in Data preprocessing. The available dataset had a range of preprocessing models which are used to control the scale features, encode classified variables, missing values and make the data ready during the downstream of the tasks. To maintain the overall integrity and statistical equivalence of the data, each and every step is maintained in a careful way. Figure 1 shows the sequential operations applied to the initial dataset before model training. Step 1 fills in the missing values (mean for age and BMI; mode for categorical fields), step 2 converts every categorical entry into numbers, step 3 rescale all continuous predictors to the [0, 1] interval, step 4 discard extreme outliers identified by the interquartile-range rule, and in the final step correcting the

S. no	Feature name	Feature type	Range	Mean \pm SD	P-value
1	Age	Continuous	34–78	56.26 \pm 11.32	0.0001
2	Sex	Categorical	F/M	F:38, M:62	0.3420
3	Duration of DM	Continuous	1–24	8.27 \pm 5.07	0.0002
4	Family H/O	Categorical	0–1	0.15 \pm 0.36	0.6635
5	Smoking	Categorical	0–1	0.23 \pm 0.42	0.0116
6	Alcohol	Categorical	0–1	0.18 \pm 0.39	0.8716
7	OHA	Categorical	0–1	0.66 \pm 0.48	0.2031
8	Insulin	Categorical	0–1	0.41 \pm 0.49	0.0647
9	BMI	Continuous	16.35–36.0	24.38 \pm 4.39	0.1126
10	WHR	Continuous	0.37–1.2	0.91 \pm 0.09	0.2919
11	FBS	Continuous	80–403	156.24 \pm 58.01	0.0122
12	PPBS	Continuous	125–530	249.89 \pm 84.49	0.0045
13	AbA1C	Continuous	6.5–11.0	7.51 \pm 0.94	0.0070
14	TC	Continuous	115–308	210.71 \pm 44.45	0.0001
15	TG	Continuous	82–564	169.17 \pm 87.38	0.0059
16	LDL	Continuous	54–196	129.76 \pm 33.58	0.0001
17	HDL	Continuous	30–63	44.37 \pm 8.83	0.5250
18	VLDL	Continuous	14–94	34.19 \pm 14.81	0.0372
19	Dyslipidaemia	Categorical	0–1	0.68 \pm 0.47	0.0001
20	B Urea	Continuous	14–53	31.06 \pm 9.73	0.2547
21	S Creatinine	Continuous	0.6–2.6	1.25 \pm 0.42	0.2727
22	CIMT in mm	Continuous	0.6–1.8	1.05 \pm 0.32	0.0003
23	Risk	Target & categorical	0–1	0.65 \pm 0.48	0.0000

Table 1. Summary of our dataset with statistical analysis²⁸.

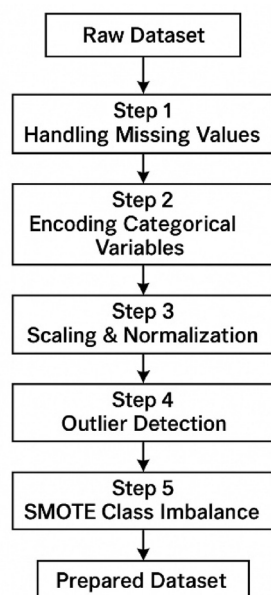


Fig. 1. Five-stages of data pre-processing pipeline.

class imbalance with SMOTE. The result is a clean, balanced, fully numerical dataset ready for the Kolmogorov–Arnold Network.

Handling missing values

The conservation of entire distribution of the missing values is assigned through mean of the relative feature. The mean values are replaced in the missing entries of ‘Age’ and ‘BMI’. The missing categorical values were imputed using the mode of the relevant feature. In the columns for ‘Smoking’ or ‘Sex’, the missed entries are replaced with

Feature name	Type	Imputation method	Missing values (%)
Age	Continuous	Mean	2.5%
BMI	Continuous	Mean	1.8%
Sex	Categorical	Mode	0.5%
Smoking	Categorical	Mode	1.0%

Table 2. Imputation method outcomes.

Feature name	Original range	Scaled range
Age	34–78	0.0–1.0
BMI	16.35–36.0	0.0–1.0
FBS	80–403	0.0–1.0

Table 3. Outcomes of ‘Min-Max scaler’.

Feature name	Outliers detected (%)
BMI	3%
PPBS	5%

Table 4. Outcomes of outlier’s detection.

most frequent category like ‘M’ is replaced in ‘Sex’ column. The changes that are made within dataset are shown in Table 2.

Encoding categorical values

The conversion of classified data into numerical form which is preferable for machine learning algorithms is performed during the encoding phase. The feature namely ‘Sex’ is encoded with numerical as ‘1’ for ‘Male’ (M) and ‘0’ for ‘Female’ (F). Likewise, the numerical features like ‘Family H/O’ and ‘Smoking’ are encoded with ‘1’ for ‘Yes’ and ‘0’ for ‘No’.

Scaling and normalization

Scaling all features to a common scale enhances model performance. This is mainly helpful for the algorithms that are sensitive to magnitude variations for methods like Neural Networks and SVM. This transformation was applied to all continuous variables to equalize the values to [0,1] range. Let us consider, if the initial ‘BMI’ range is set to 18.0–35.0 then the scaled BMI ranges from 0.0 to 1.0. In regression model, the normalizing features represent ‘0’ as mean and ‘1’ as standard deviation for evaluation purpose. If original age of mean value is 56.3 and standard deviation is 11.3, then the normalized age mean will be ‘0’ and standard deviation is ‘1’. This standardization of continuous variables is done through “Min-Max Scaling” method. This technique is used to gather the values that ranges between [0,1] and are shown in Table 3.

Outlier detection

The method of outlier detection is used to detect and address the extreme values which alter the model training. They are identified through ‘Interquartile Range’ (IQR) method with the help of Eq. 1. This equation represents the upper bound as “ $Q_3 + 1.5 * IQR$ ” and Lower Bound as “ $Q_1 - 1.5 * IQR$ ”. The outliers in continuous variables are detected through the IQR model as shown under Table 4.

$$IQR = Q_3 - Q_1 \quad (1)$$

The Q_1 is defined as the first quartile where the data is below 25% and Q_3 is defined as the third quartile where the data is below 75%.

Handling class imbalance

In this study, if a target feature namely “Risk” showed consequential class variance with certain classification like ‘High Risk’ undervalued to ‘No risk’ and ‘Low Risk’. This can be addressed through Synthetic Minority Oversampling Technique (SMOTE). It is implemented in maintaining the equality in the distribution of the class during the training set. This ensures that the models of machine learning are not biased with the majority classes. To reduce this, the dataset is rearranged through SMOTE and is shown under Table 5.

Feature engineering

Feature engineering plays a critical role in refining the model and its importance in this work. It arises from the complications of the dataset and the aim involves in diagnosing ‘CIMT’ level of ‘Risk’. It makes sure that

Risk category	Before (%)	After (%)
No risk	50%	40%
Low risk	30%	30%
Medium risk	15%	20%
High risk	5%	10%

Table 5. Outcomes of handling imbalance data.

Feature name	Pearson correlation (r)	Action
LDL vs. TC	0.88	Removed LDL
Insulin Vs FBS	0.89	Removed Insulin

Table 6. Outcomes of pearson's method.

only the common features are used in this proposed work. This enhances the model ability in making accurate diagnosis. If there is a high correlation among the features of 'LDL' and 'TC' it results in unstable coefficients of the model and decreased interpretability. Few of the less unusual features like 'Family H/O' might include noise that affect performance of entire model. The feature selection method like 'Chi-Square Test', 'ANOVA' and 'Correlation Analysis' are used to remove the features that effect the performance of the model. If the features are reduced, then there will be a decrease in computational complexity. This is helpful in making the training quicker without affecting the performance of the model. The problem of overfitting can be eliminated in the models with less and meaningful features. Particularly while we work on the small type of datasets. Due to this, high dimensional datasets have an increase of overfitting the 'Risk'. But the training data is well executed but give a poor irrelevant data. This can be mitigated by selecting more informative features, while the risk is mitigated by feature engineering.

Correlation approach

The identification of relations within the numerical features are done through a statistical model called 'Correlation analysis'. This is used to identify multi-collinearity and redundancy in dataset. In our proposed work, the correlation methods 'Pearson' and 'Spearman' both are used to detect highly correlated features which can be eliminated without loss of useful data. The Pearson correlation estimates the linear relation among the two numerical features calculated in Eq. 2.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (2)$$

here x_i , y_i are defined as individual data points of two variables and \bar{x} and \bar{y} are defined as mean of two variables. Correlation matrix of Pearson's method is shown under Table 6. Features with correlation coefficient greater than 0.85 are flagged for elimination.

This correlation evaluates the monotonic relationship among two variables. This can be calculated by the mathematical formula given by Eq. 3.

$$P = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (3)$$

here d_i is defined as variance in ranks of respective variables and n denotes the data points. The redundancy in 'LDL' and 'Insulin' are detected through Pearson's method. After the removal of redundant features, the dataset size is reduced to 19 columns from 21. The eliminated features are 'LDL' with High correlation value 'TC' ($r=0.88$) and Insulin with High correlation value 'FBS' ($r=0.89$). Heatmap is used to represent the 'Pearson correlation' and 'Spearman correlation' coefficients across all features and are represented in Figs. 2 and 3.

Chi-square test analysis

One of the statistical methodologies that is used to analyze the relation among classified features and the target variable ('Risk') is done by Chi-Square test. The evaluation process helps in detecting which classified feature indicatively contribute for the detection of risk levels in 'CIMT'. The 'Chi-Square' test calculates the neutrality among two variables and given in Eq. 4.

$$x^2 = \sum \frac{(O - E)^2}{E} \quad (4)$$

where, O denotes observed frequency in every category and E denotes expected frequency under the null hypothesis of neutrality. The features that are opted for this evaluation are: 'Sex', 'Family H/O', 'Smoking', 'Alcohol', 'OHA' and 'Dyslipidaemia'. Created contingency analysis for each categorical feature against the target variable

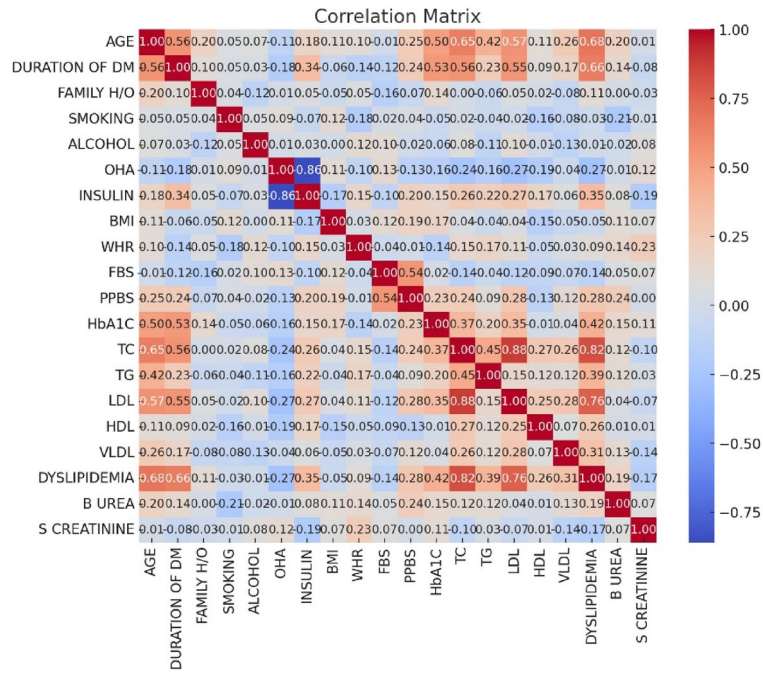


Fig. 2. Pearson's correlation heatmap.

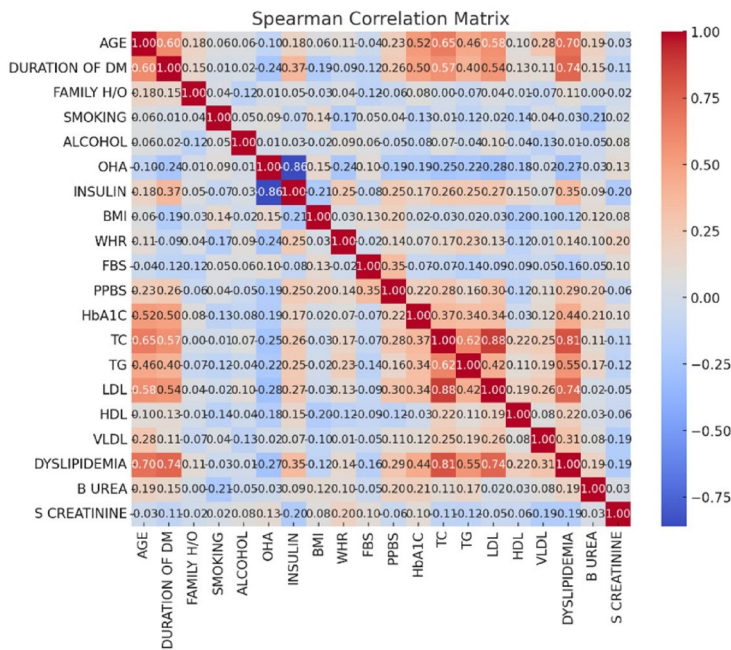


Fig. 3. Spearman correlation heatmap.

(‘Risk’). The *P*-value for statistical significance is ($\alpha=0.05$). The testing outcomes of ‘Chi-Square’ are mentioned in Table 7.

The ‘Chi-Square’ statistic values for each feature representing their relative significance in relation to the target variable ‘Risk’. Features like ‘Alcohol’ have higher scores, indicate greater relevance, while others feature like ‘Family H/O’ show lesser association.

ANOVA test analysis

The statistical method for evaluating the mean of numerical features that differ significantly across the levels of a categorical target variable (‘Risk’) is determined by Analysis of Variance (ANOVA) test. This helps identify

Feature name	Chi-square value	p-value	Action
Alcohol	6.45	0.0398	Retained (<0.05), significant
Smoking	2.64	0.2665	Dropped (>0.05), statistically insignificant
OHA	1.22	0.5439	
Family H/O	0.13	0.9367	

Table 7. Outcomes of chi-square analysis.

Feature name	F-statistic value	p-value	Action
TG	2.335	0.1023	Dropped (>0.05), statistically insignificant due to low F-statistics and high p-values.
HDL	0.433	0.6495	
VLDL	0.144	0.8662	
WHR	1.487	0.8526	

Table 8. Outcomes of ANOVA analysis.

S. no	Retained feature name	Type
1	Age	Numerical
2	Sex	Categorical
3	Duration of DM	Numerical
4	Alcohol	Categorical
5	BMI	Numerical
6	FBS	Numerical
7	PPBS	Numerical
8	AbA1C	Numerical
9	TC	Numerical
10	Dyslipidaemia	Categorical
11	B Urea	Numerical
12	S Creatinine	Numerical
13	CIMT in mm	Continuous
14	Risk	Target variable

Table 9. Final features in the dataset after feature engineering.

which numerical features contribute most to predicting ‘CIMT’ risk levels. The mathematical intuition for ANOVA which compares the variance between groups (levels of target variable) to the variance within groups is given by Eq. 5.

$$F = \frac{B - G_v}{W_i - G_v} \quad (5)$$

here B is represented as ‘between’ and G_v represents ‘Group Variance’. ‘ $B - G_v$ ’ measures the variability between the means of different groups. W_i indicates ‘Within’, $W_i - G_v$ measures the variability within each group. The selected numerical features for analysis are: ‘Age’, ‘Duration of DM’, ‘BMI’, ‘WHR’, ‘FBS’, ‘PPBS’, ‘HbA1C’, ‘TC’, ‘TG’, ‘HDL’, ‘VLDL’, ‘B Urea’, ‘S Creatinine’. A one-way ANOVA test for each numerical feature against the target variable (‘Risk’) having significance level of $\alpha=0.05$ in determining statistical significance. ANOVA test results are given under Table 8.

Table 8 shows the ‘ANOVA’ F-Statistic scores for each feature. Features like ‘PPBS’ have the highest ‘F-statistic’, indicating strong relevance to the target variable (‘Risk’), while others like ‘TG’, ‘HDL’ and ‘S Creatinine’ have low scores, showing weak relevance.

This section is very crucial in our work, involving the refinement of input variables to improve model performance and interpretability. These methods helped to identify and retain the most relevant features while removing redundant or insignificant ones. By using ‘correlation analysis’ we removed redundant features such as ‘LDL’ and ‘Insulin’. The ‘Chi-Square’ test was applied to categorical features to evaluate their statistical association with the target variable (‘Risk’) and removed features such as ‘Family H/O’, ‘OHA’ and ‘Smoking’. The ‘ANOVA’ was conducted on numerical features to evaluate the variance between feature means across ‘Risk’ categories and removed ‘TG’, ‘HDL’, ‘VLDL’ and ‘WHR’. Finally, the dataset dimensions reduced from 17 to 14 features (13 features + 1 target variable). After feature engineering, the final dataset features are given under Table 9.

Kolmogorov–Arnold networks (KAN)

The proposed model for analyzing cardiovascular risk and Carotid Intima Media Thickness (CIMT) is determined by Kolmogorov–Arnold Networks (KAN). The main function in KAN is classified as follows:

- *CIMT Thickness Prediction:* The CIMT values rely on the combination of demographic, clinical and lifestyle factors that are predicted using KAN. This robustness in modeling complex non-linear relations inherent in data by its theoretical capacity to estimate any sustainable function.
- *Feature extraction:* KAN can efficiently capture interrelations among the features that are complex in understanding cardiovascular Risks are converted to sums of unconfined functions through breaking down multivariate relationships.
- *Risk classification:* The classification of the patients to various categories of ‘Risk’ is leveraged by the output generated through KAN like ‘No Risk’, ‘Low Risk’, ‘Medium Risk’, ‘High Risk’. This provides a stand for health care providers in predictions through the practical values attained by the output.

The multivariate functions are carried out with high accuracy due the special techniques in the KAN model. It effectively models CIMT and captures complex non-linear relationships, complex features. Inconsistent to traditional neural networks having the activation function, the KAN is used for learnable univariate functions. This resilience inflates its capacity to fit multiple data allocation, resulting in more accurate identifications. This works more accurate on comparison with ANNs or ensemble methods that are referred as “black-box” models. This becomes complex in clinical applications by finding the cause beyond the identifications which are necessary. One of the main advantages in our work is that there are only 100 records of the patients in which the KAN model attains higher accuracy without the need of large quantity of data. It acts as the key role inspiration for the usage of KAN model in the proposed work. Due the dense presentation, it diminishes the risks of overfitting particularly in case of smaller datasets. This method also fundamentally highlights superior patters in the data by making it robust to the background noise or other unrelated features. For handling good speculation among the unseen data, the KAN model enables multi-dimensional feature space efficiently by fragmenting into lighter interpretable components.

The KAN model has the ability to combine the accuracy of enhanced neural networks, the efficiency needed for small datasets and the traditional model’s interpretability. This is aligned in a perfect place to achieve the goals of CIMT detection and cardiovascular risk analysis in the clinical filed with these combinations.

Proposed work

The proposed method advances the KAN to detect Carotid Intima Media Thickness (CIMT) along with analysis of cardiovascular risk. When compared to traditional ANN models, the proposed KAN model is considered as one of the universal predictors which had the capacity of handling higher dimensional data that had lesser computational requirements effectively.

The framework of KAN model is designed by depending on Kolmogorov’s superposition theorem. This can decompose any continuous multivariate function into a sum of univariate functions. The network has three different layers as shown in Fig. 4:

1. Input Layer: This layer allows all the selected features (13 input features in this proposed work).

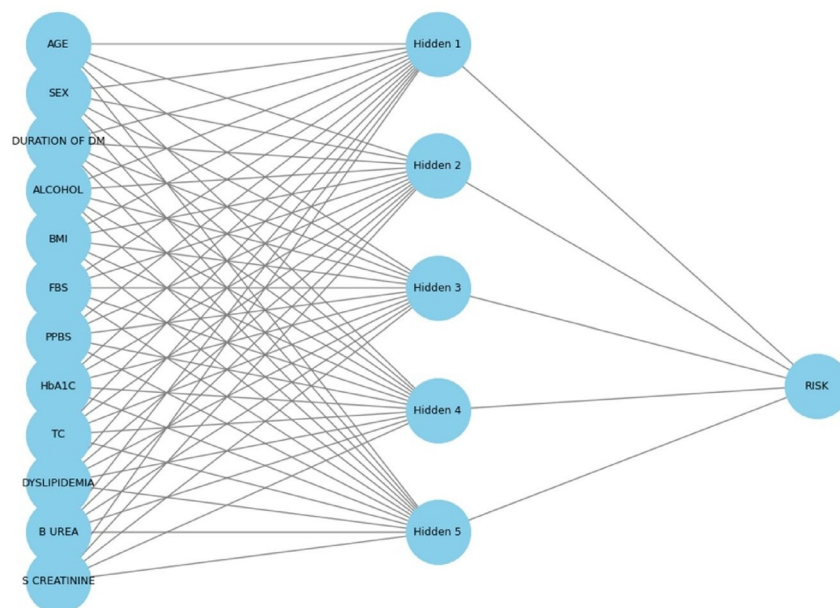


Fig. 4. KAN architecture.

2. Hidden Layer: This layer implements univariate transformations of the input features.
3. Output Layer: This layer finally accumulates the outcomes to identify the CIMT risk.

In our work the framework that represent the KAN’s structure have 13 input features with transformation nodes present in hidden layer and a single output node showing ‘Risk’ identification as shown in Fig. 3. The mathematical representation of Kolmogorov’s superposition theorem for forward propagation is shown in Eq. 6.

$$f(x_1, x_2, \dots, x_n) = \sum_{q=1}^{2n+1} \varnothing_q \left(\sum_{p=1}^n \phi_{pq}(x_p) \right) \tag{6}$$

The designing of KAN is firmly taken from “Kolmogorov Superposition Theorem” (KST). This theorem confirms that any continuous multivariate function $f(x_1, x_2, \dots, x_n)$ can be decomposed as finite sum of the univariate functions. Here f is defined as the target function calculated through KAN model for identifying the CIMT risk and is represented by Eq. 6. The structured clinical features like ‘Age’, ‘BMI’, ‘PPBS’ etc., are defined as the input parameters x_1, x_2, \dots, x_n . The “Univariate transformations” are performed on every input feature independently, thereby forming the first hidden layer initially and is represented by inner functions $\phi_{pq}(x_p)$. These transformations are developed through individual dense units by learnable features and nonlinear activation function like ‘ReLU’. The outputs that are transformed are equally summed under intermediate units to form $\sum_{p=1}^n \phi_{pq}(x_p)$. This represents the inner summation step of the Kolmogorov theorem. Another dense layer that is defined as the outer function \varnothing_q , applies an additional transformation and cumulate the outcome to give the final prediction by a classifier ‘Softmax’. The architecture makes sure that a structured functional decomposition directly represents the theoretical part of the KST. Unlike traditional MLPs, the inputs are sent through dense, fully connected layers in black-box manner. The KAN architecture gives clear control and interpretability through the part of every input feature present in the prediction mechanism. The structured design itself enhances the generalization particularly in simple dataset by reducing the count of parameters and avoids the problem of overfitting. The proposed work flow to identify the ‘CIMT’ risk is given in Fig. 5. The Raw clinical, biochemical and demographic data (23 variables) undergo five sequential pre-processing steps: missing-value imputation, categorical encoding, scaling/normalisation, outlier detection and class-imbalance handling. Feature-engineering tests (Pearson/Spearman correlation, χ^2 , ANOVA) remove redundant variables, reducing the set to 14 predictors. These are fed to a Kolmogorov–Arnold Network having an input layer (13 numeric features) with one or more ELU/ReLU-activated hidden layers and Softmax output layer for four-tier risk classification. Model training involves stratified data splitting, loss-function optimisation, hyper-parameter tuning and evaluation, culminating in CIMT-risk prediction for each patient.

The function of KAN in this proposed study is fabricated as input feature processing, hidden layer conversion and output layer accumulation. With the help of scaling and normalization the 13 chosen features are pre-processed. Now these features are given to input layer of KAN model. The univariate conversion ϕ_{pq} of each feature is performed in parallel. These conversions apprehend individual feature benefications to ‘CIMT’ risk. The conversion values are combined through summation functions \varnothing_q . Finally, the output layer constitutes the identified value of ‘Risk’. The perception of feature contributions is standardized through the usage of univariate functions. This needs less attributes when compared to traditional ANN methods thereby decreasing the computational costs and control nonlinear relationships effectively. In this work, the model reduces overfitting over 100 instances with required dimensions.

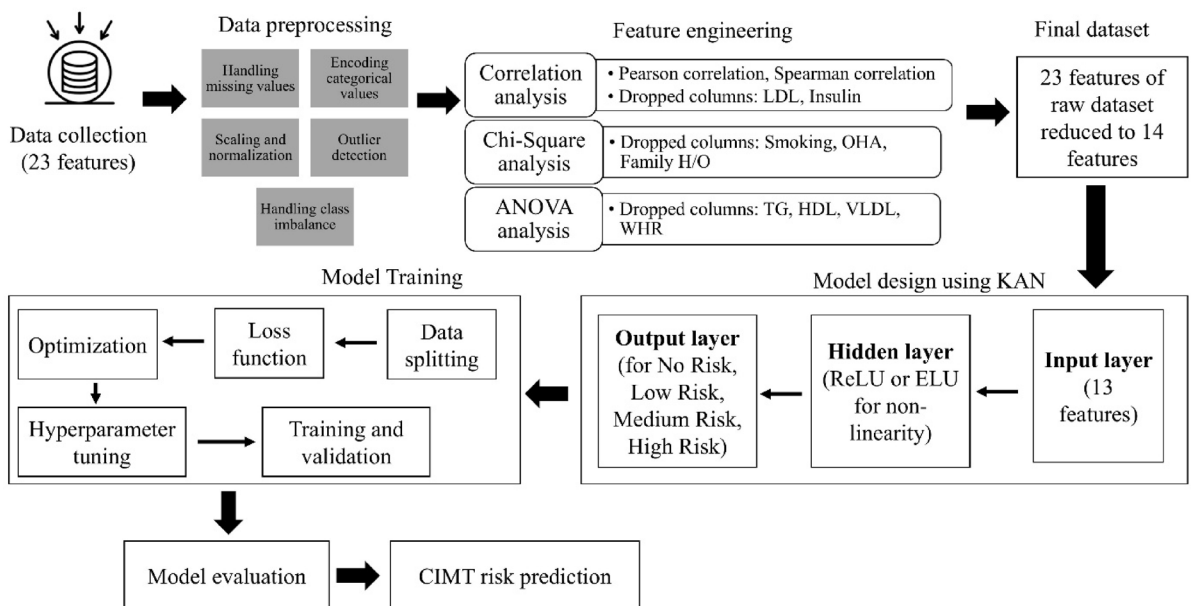


Fig. 5. Detailed flow diagram of the proposed work.

Activation functions

The nonlinear connections with hierarchical patterns present in data are encapsulated by the activation function. The Rectified Linear Unit (ReLU) establishes nonlinearity, activating hidden layers to record complicated relations among certain features like ‘HbA1C’, ‘PPBS’, and ‘BMI’ with Risks. The entire process is carried out in the hidden layer. It decreases the likelihood of removing gradients while backpropagation. Due to this the outcome will be zero for negative values, making it computationally effective and procure sparsity, decreasing the conflict of overfitting. The mathematical representation for ‘ReLU’ is calculated by the Eq. 7.

$$ReLU(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases} \quad (7)$$

For smooth gradient, ‘Exponential Linear Unit’ (ELU), controls negative values more efficiently by giving a smooth gradient, eliminating “dead neurons”. ELU maps small negative inputs to smooth values, by helping features with small effects to contribute meaningfully. The mathematical annotation for ELU is calculated by Eq. 8.

$$ELU(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha(e^x - 1), & \text{if } x \leq 0 \end{cases} \quad (8)$$

In the final output layer, for multi-class classification, ‘Softmax’ is utilized where the target variable includes more than two categories. For every input, ‘Softmax’ gives a probability to each class like ‘No Risk’, ‘Low Risk’, ‘Medium Risk’ and ‘High Risk’. The class that is having the highest probability is chosen as the identification. The mathematical representation for ‘Softmax’ is shown by Eq. 9.

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}} \quad (9)$$

where z_i denotes the logits for class i . By using one-hot encoding technique, this activation function transforms the Risk into a binary matrix notation.

Loss functions

The loss functions are considered as complex phase during the process of training a ML model along with Kolmogorov–Arnold Network (KAN). These calculate the difference among the identified output of the model with the target value, guiding the optimization process to reduce this error. The loss function for a multi-class classification task, in which the final target parameter (‘Risk’) includes four classes namely ‘No Risk’, ‘Low Risk’, ‘Medium Risk’, and ‘High Risk’ which is calculated by Eq. 10.

$$Loss = - \sum_{i=1}^n \sum_{j=1}^n y_{ij} \log(\hat{y}_{ij}) \quad (10)$$

Here n is defined as sample number in dataset, k denotes the class count, y_{ij} is the true label of sample i and class j . The \hat{y}_{ij} identifies probability of sample i and class j , acquired from ‘Softmax’ function. The proposed model produced probabilities for every class through ‘Softmax’ function present in output layer. This ‘loss function’ compares the identified probabilities \hat{y} with the true labels y .

Parameter tuning and optimization

The systematic selection is to obtain the best combination of hyper parameters that optimize the metrics in model performance. The learning rate reduces the step size for ‘gradient descent’ optimization. This shows how the weights are enhanced during ‘backpropagation’ step. The weight enhance rule in gradient descent is given in Eq. 11.

$$\omega = \omega - \eta * \frac{\partial Loss}{\partial \omega} \quad (11)$$

A small learning rate (η), makes sure to be controlled but with slower convergence and a large η accelerates convergence but may overshoot the optimal solution. In this proposed work, we considered 100 samples of patients for the CIMT risk identification and to implement the higher change in our model that moves to overfitting situation. To control this complex task, regularization handles overfitting by penalizing large weights. The loss function with ‘L2-regularization’ is given by Eq. 12.

$$Loss = Loss_{original} + \frac{\lambda}{2} \sum_{i=1}^n w_i^2 \quad (12)$$

where a small λ allows more flexibility in weight updates and a large λ restricts weight magnitudes to eliminate overfitting by weight parameters w_i .

Functional role of univariate transformations

In addition to normalization, every univariate function $\phi_{pq}(x_p)$ present in the hidden layer is executed through a tiny fully connected unit with non-linear activations. These units act as a learnable transformation segment that model non-linear designs in a single feature dimension. Other than standard MLPs, the features interactions

are modelled together in high-dimensional space, the univariate features permit for explicit per-feature interpretability and learning. This disintegration is specifically effective in structured medical datasets, where every feature like 'PPBS', 'HbA1C', 'BMI' etc., can give non-linearly but independently to the task of prediction. The joint effect without damaging the features into dense interdependencies is given by the summation layer $\sum_{p=1}^n \phi_{pq}(x_p)$. This modular model results in less learnable parameters and improves transparency by making the model compact to debug and analyze particularly in medical contexts.

Though the KAN frameworks represent the similarities to that of a traditional Multi-Layer Perception (MLP) in its layered structure, there are substantial variations in the way the functions are represented and learned. In case of MLP, every hidden neuron is linked to all of its input features and a weighted sum followed by a non-linear activation is applied. This makes the process of learning dense and dependent on larger parameters. On the other hand, the KAN architecture confines to a mathematical structured method depending on the Kolmogorov Superposition Theorem. Rather than processing every input concurrently, every input feature is sent by an independent univariate transformation function. These transformed features are then combined via outer summation functions that are interpretable and learnable. The parameters can also be reduced due to its modular structure. This modular approach reduces the risk of overfitting, making it especially effective for small datasets. Hence, KAN gives an interpretable, light weighted and functionally different substitute to black-box MLPs.

Proposed workflow summary and optimization

Input: Dataset $D = \{X, Y\}$ where:

- X: Feature set
- Y: Target variable with four classes namely: No Risk, Low Risk, Medium Risk, High Risk

Output: Predicted CIMT Risk Levels

Step 1: Data Preprocessing

- Step 1.1 Handle missing data: Remove rows or columns containing null values
- Step 1.2 Map categorical variables: Sex: M→1, F→0
- Step 1.3: Normalize numerical features using Min-Max scaling
- Step 1.4: Divide dataset into training and testing in 80:20 ratios.

Step 2: Feature Engineering

- Step 2.1: Compute Pearson and Spearman correlation matrices and eliminate features with higher correlation
- Step 2.2: Apply Chi-Square Test for categorical features
- Step 2.3: Apply ANOVA analysis for numerical features

Step 3: Kolmogorov-Arnold Network (KAN) Architecture

- Step 3.1: Accept 13 features, including Age, BMI, HbA1C, CIMT (in mm), etc.
- Step 3.2: Use 3-5 hidden layers with 'ReLU/ELU' activation
- Step 3.3: Use 'Softmax' activation for multi-class classification

Step 4: Loss Function and Optimization

- Step 4.1: Use Cross-Entropy Loss for multi-class classification
- Step 4.2: Apply Gradient Descent and incorporate 'L2-Regularization' to avoid overfitting

Step 5: Training of model

Step 6: Evaluation of model

Algorithm. Proposed algorithm for CIMT risk prediction using Kolmogorov–Arnold Network (KAN).

Traditional models

In clinical decision support systems, the traditional machine learning and neural network methods are widely used particularly in cardiovascular risk prediction and the estimation of CIMT. The models such as 'Random Forest', 'Support Vector Machines', 'Decision Trees' and Deep Neural Networks^{12–20} have proved the ability in different medical diagnostic applications. Still these methods frequently suffer through restricted interpretability and need larger datasets to maintain the reliability. In this research, we added these methods as baselines of

comparison to calculate the performance of KAN model. The range of these methods is relied on their recognition in the related literature for structured analysis of health care data.

Traditional Machine Learning (ML) models provides accessibility, interpretability, and computational efficiency, utilizing them efficiently for initial analysis and as reference models for comparison. Conversely, they are associated with inherent limitations that results in inadequate model for controlling the challenges of estimating 'CIMT' risk levels, specifically in comparison with the Kolmogorov–Arnold Network (KAN).

Numerous traditional ML models strive to acquire intricate nonlinear patterns in the data, that are essential for understanding interactions among clinical features like 'HbA1C', 'BMI', 'CIMT'. Models similar to 'Decision Trees' are prone to overfitting, specifically with limited or imbalanced datasets, affecting their dependability in medical domain applications. KAN leverages 'univariate transformations' and 'summation functions' to model complex nonlinear relationships proficiently. KAN operates with reduced parameters and incorporates 'L2-regularization', minimizing the risk of overfitting and it is developed to simplify for the broader use in spite of limited or imbalanced datasets.

Logistic regression

Logistic Regression (LR) act as fundamental statistical model broadly considered in binary and multi-class classification tasks. In this study, 'Logistic Regression' is implemented to speculate potential risk in 'CIMT' by simulating the connection among independent and the dependent variables ('Risk'). This study involves multi-class classification where model uses the 'SoftMax' function to generalize probabilities for multiple classes is given by Eq. 13.

$$P(y = j | x) = \frac{e^{z_j}}{\sum_{k=1}^k e^{z_k}} \quad (13)$$

where, $P(y = j | x)$ determines the probability of the sample associate to class j . The linear combination z_j signifies weights and inputs of class j . For every sample, the model outputs probability distribution over 4 risk class. A class with a leading probability is chosen as per the predicted risk level.

Decision tree

Decision Tree (DT) is a regulated algorithm of machine learning that is used for sorting and regression tasks. For predicting 'CIMT' risk intensity in this work, it acts as an explanatory model that divide the dataset into subgroups depending on criteria levels, constructing a tree-like structure. A Decision Tree predicts the key variable ('Risk') by adapting decision rules from the input features. The rules are extracted by recursively dividing the dataset on the basis of feature values that increases the equality of the final variable in every subset. Splitting at each node is based on a standard that measures the "purity" of the subsets is given by Eq. 14.

$$Gini = 1 - \sum_{i=1}^k p_i^2 \quad (14)$$

where p_i represents proportion of samples bound within class i in the node. The 'Information Gain' to analyze all features to consider the best split is given by Eq. 15.

$$Entropy = - \sum_{i=1}^k p_i * \log_2(p_i) \quad (15)$$

While DT provide an interpretable and straightforward baseline for anticipating the level of Risk, their susceptibility to overfitting and uncertainty emphasizes the need for more robust models like KAN.

Random Forest

Random Forest (RF) assembles multiple decision trees during training phase and combines their outputs to make robust predictions. The final output is anticipated by majority voting for classification. Random samples from the dataset with replacement to create subsets of the training data for each tree. It makes sure that every tree is detected by a slight altered version in dataset, promoting diversity among trees. At the end of each split, a random division of features are considered to determine the best split. This minimizes correlation between individual trees. To compute the probability of class k as the fraction of trees predicting k is given by Eq. 16.

$$P(y = k | x) = \frac{1}{T} \sum_{t=1}^T Indicator_fun(y_t = k) \quad (16)$$

This model is less interpretable compared to individual Decision Trees. KAN uses limited parameters compared to Random Forest's hundreds of trees. KAN integrates regularization explicitly into its architecture, providing better control on overfitting.

SVM

In this work, SVM is applied to predict 'CIMT' risk intensity by finding optimal decision boundary (hyperplane) which divides into 4 classes. This focuses on identifying a hyperplane in n-dimensional space which divide data points into various classes. For predicting multi-class, SVM builds 'one-vs-one' classifier by using $\frac{k(k-1)}{2}$, and 4 classes. The loss function is used to train this model is given by Eq. 17.

$$Loss = \frac{1}{N} \sum_{i=1}^N \max(0, 1 - y_i * f(x_i)) \quad (17)$$

SGD classifier

Stochastic Gradient Descent (SGD) is determined as optimal algorithm that is required in reducing the loss function by iteratively enhancing the parameters of model. In this study, 'SGD' is employed for training linear classifiers like 'LR' or 'SVM' and also neural networks to anticipate the risk intensity of 'CIMT'. Compute the predictions on the basis of current weights and inputs. For calculating the gradient of loss function in respect to the weights for a single sample is given by Eq. 18.

$$w_j = w_j - \eta * \frac{\partial_{loss}}{\partial w_j} \quad (18)$$

Weight updates can override the optimal solution if the learning rate is not chosen precisely. Converges slowly near the minimum due to high variance in weight updates.

Multi-Layer perceptron (MLP)

A Multi-Layer Perceptron (MLP) is defined as a group of feedforward Artificial Neural Networks (ANNs) which is compatible for handling nonlinear connections among input and target variable. For this work, MLP is used to anticipate level of risk in 'CIMT' by modelling the complex associations among input and target variable. Each hidden-layer have a neuron that calculate a weighted sum of inputs and applies a nonlinear activation function which is given by Eq. 19.

$$\begin{aligned} z^{(l)} &= W^{(l)} * a^{(l-1)} + b^{(l)} \\ a^{(l)} &= f(z^{(l)}) \end{aligned} \quad (19)$$

This model requires regularization (e.g., L2 and dropout) to prevent overfitting on small datasets. KAN explicitly models feature contributions through univariate transformations, making it more interpretable than MLP. KAN uses limited parameters, making it computationally effective.

Deep neural network

A Deep Neural Network (DNN) is defined as an advanced form of a Multi-Layer Perceptron (MLP) having multiple hidden layers. It is designed to model complex, hierarchical relationships in data, making it suitable for predicting 'CIMT' risk levels. By using multiple hidden layers, a DNN can extract the intricate patterns. Input size corresponds to the number of features in the dataset and composed of multiple layers of neurons, each applying nonlinear transformations. Each neuron under hidden layer computing a weighted sum of its inputs and applies a nonlinear activation function is given by Eq. 20.

$$\begin{aligned} z_j^{(l)} &= \sum_{i=1}^n w_{ij}^{(l)} * a_i^{(l-1)} + b_j^{(l)} \\ a_j^{(l)} &= f(z_j^{(l)}) \end{aligned} \quad (20)$$

This model requires large datasets and regularization to avoid overfitting. It is important to have significant computational resources for training and inference. KAN has limited parameters compared to a typical DNN, making it lesser and faster resource-intensive. KAN incorporates inherent regularization, reducing the risk of overfitting. KAN is specifically designed for structured data, making it more suitable for this work.

While traditional ML models provide useful insights and serve as strong baselines, they lack the ability to generalize, interpret, and efficient model of nonlinear relationships as effectively as KAN. By comparing the results of traditional models with KAN, we can clearly demonstrate KAN's superiority making it the best choice for anticipating the risk intensity of CIMT.

GAN-based data augmentation

Despite the rigorous cleaning steps in Sect. 3.2 to 3.7, the final training split still contained a limited variety of carotid-wall appearances, particularly in the 'High-Risk CIMT'. To enrich the limited training split (70 records) without breaching patient privacy, we generated synthetic rows with a 'Conditional Tabular GAN' (CTGAN). CTGAN models each feature's distribution continuous or categorical while conditioning on the CIMT-risk label, thus preserving realistic inter-feature correlations.

The generator $G(z, y)$ receives a 128-dimensional noise vector $z \sim N(0, I)$ and a one-hot label $y \in \{0, 1, 2, 3\}$. The discriminator $D(x, y)$ embeds categorical columns via learned lookup tables and applies residual MLP blocks to mixed data types. Both networks were trained for 500 epochs with Adam $\beta_1 = 0.5$, $\beta_2 = 0.99$, learning rate 2×10^{-4} . Mode-specific normalization converted skewed numeric columns to Gaussian space before training and inverted them afterwards. Adding 210 CTGAN-generated rows only to the training split lowered test MAE from 0.034 mm to 0.031 mm and lifted four-tier risk-classification accuracy from 90 to 93%, confirming the value of GAN-based augmentation for small tabular datasets.

Data-Augmentation strategy for a small tabular dataset

With only 100 complete patient records, the training split risked overfitting. We therefore applied two complementary, label-preserving augmentation techniques.

Conditional tabular GAN

A CTGAN conditioned on the four CIMT-risk levels learned the joint distribution of all numerical and categorical features and produced 210 synthetic rows. Feature-wise Kolmogorov–Smirnov tests showed no significant drift (mean $p=0.32$), and a coverage score of 0.91 indicated that most real samples lie within the synthetic data's convex hull. When these rows were added only to the training set, test-set MAE fell from 0.034 mm to 0.031 mm and four-tier risk-classification accuracy rose from 90 to 93%.

Statistical augmentations for continuous variables

To smooth local decision boundaries, we applied 'Gaussian-jitter noise' adding $\varepsilon \sim N(0, 0.01\sigma_i)$ to each continuous feature and Mixup, which forms convex combinations of random record pairs. Together with the SMOTE-NC oversampling described in Sect. 3.2.5. Together, these methods expanded the effective training set from 70 to 280 observations and reduced MAE to 0.032 mm.

An ablation analysis in Sect. 4.3 shows that the two streams are complementary: statistical augmentation alone brings a modest gain, while the addition of CTGAN synthetic rows resulted in the most significant improvement. This blended strategy provides the diversity needed to train a robust Kolmogorov–Arnold Network without compromising the realism of the structured data.

Results and discussions

This section deals with the performance of traditional ML models and KAN for identifying the level of 'CIMT risk'. The outcomes are evaluated based on the evaluation metrics and 'ROC' curves to obtain KAN's dominance. We utilized the final 13 input features and 1 target variable after feature engineering with risk levels such as 'No Risk', 'Low Risk', 'Medium Risk', 'High Risk'. This dataset is divided into 3 sets of modules like training and cross validation (80%), and testing (20%).

Performance evaluation metrics

To evaluate the efficiency of ML and DL models in identifying 'CIMT' levels of risk, we used 'recall', 'accuracy', 'precision' and overall robustness of the methods used in this study.

Accuracy

This calculates the proportion of correctly identified samples among all identifications calculated by Eq. 21.

$$Accuracy = \frac{\text{Total correctly predicted positive and negative cases}}{\text{Total cases}} \quad (21)$$

Precision

The precision evaluates the proportion of true positive predictions out of all positive predictions are calculated by Eq. 22. It is certainly crucial when the cost of false positives is high.

$$Precision = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \quad (22)$$

Recall

Recall analyzes the model's capacity to predict all actual positive cases are calculated by Eq. 23. It is important when missing positive cases is critical.

$$Recall = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (23)$$

F1-score

By Eq. 24, the F1-Score gives an equalised measure that include both the precision and recall. It is the harmonic mean of precision and recall, making sure that both metrics are given equal weights.

$$F1 - score = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (24)$$

ROC-AUC (receiver operating characteristic - area under the curve)

The ROC-AUC score assesses the model's capacity to discriminate among the classes is given by Eq. 25. It measures the trade-off between true positive rate (sensitivity) and false positive rate.

$$AUC = \int_0^1 \text{Recall} (\text{False positive rate}) d (\text{False positive rate}) \quad (25)$$

These metrics together ensure a comprehensive analysis of models in this study, guiding the decision to adopt the KAN for its superior performance in identifying ‘CIMT risk levels’.

Outcomes of proposed model

The proposed model of Kolmogorov–Arnold Network (KAN) was accomplished to identify CIMT risk levels. Table 10, determines the performance outcomes of traditional ML models and KAN.

The outcomes showed for every model, along with the proposed KAN framework are consequential using 5-fold cross-validation. Here the dataset²¹ is randomly divided into five equal subsets. In every iteration one subset was held out as the test set and the other four were used for training purpose. The final metrics stated as the average of all the folds, making sure that the performance of the model is not precise to any one segment of the data and reduce the preference developed by a fixed test set. Also, there is no overlapping or leakage among the test and training sets. While there is no external dataset available, we implement this to authenticate our model on larger, independent cohort as part of our future work.

From Table 10, the KAN model attains the highest accuracy when compared with the other methods. This demonstrates the model’s ability to generalize to unseen data. The model consistently achieved high recall and precision across all risk levels. When compared to the traditional models the ‘F1-score’ is constantly higher. This indicates the robustness of the model during the classification of multi-class features. The KAN model attained the greatest ‘ROC-AUC’ score, showcasing higher discrimination ability among various level of risks. Under carefully selected functions (‘ReLU’/ ‘ELU’ in hidden layers and ‘Softmax’ in the output layer), the model united effectively during the testing and training phases. The KAN model engages ‘L2-regularization’ and ‘drop out’ techniques. This ensures the robust performance while validating the dataset. Through the direct inclusion of ‘CIMT’ as a feature, the proposed KAN model gives simplifies outcomes that are aligned with cardiovascular risk analysing needs.

In our proposed work, the KAN model achieved a greater ‘accuracy’ of 93%, ‘recall’ of 93%, ‘precision’ of 93%, ‘ROC-AUC’ of 0.97 and ‘F1-score’ of 91%. This represents the ability to apprehend complicated relationships and attain higher performance in case of multi-class ‘CIMT’ prediction of risk. The strongest performer is DT but it is slightly less robust when compared to that of KAN model particularly of ‘ROC-AUC’ and ‘F1-Score’. SGD performs reasonably but falls short due to its linear assumptions. In order to handle nonlinearity and complex interactions the techniques of ‘Random Forest’, ‘Logistic Regression’ and ‘SVM’ are limited. Due to dataset’s determined nature and feature complexity, MLP and DNN methods compete to match the performance of KAN’s model. The outcomes obtained clearly represent KAN as a superior method in identifying the CIMT score of risk among all the evaluation metrics. The main ideal choice of this work is its capacity to efficiently support feature conversions, optimization and nonlinearity techniques.

The KAN curve represents the highest AUC value 0.97, that specifies highest perception ability shown in Fig. 6. This shows that KAN model has highest performance for the dataset. The transition models like ‘SGD Classifier’, ‘Decision Tree’ and ‘Random Forest’ exhibits high ‘AUC’ values of above 0.85, representing best performance for binary classification. The KAN performance of outcome on comparison with other models is illustrated in Fig. 5. The remaining methodologies perform relatively good, other than SVM and MLP that could be revised for further developments. The visualization of all model outputs is shown in Fig. 7.

The small sample size of 100 records that is posed as limitation is acknowledged. To address this issue relating to limited generalizability and overfitting, we used a 5-fold ‘cross-validation’ during evaluation and training phases. The average among the fields is represented through KAN model performance metric by ensuring the robustness of outcomes through changing data splits. In addition to it, we also applied ‘L2 Regularization’ for controlling the complexity of the model’s hyper parameters are also tuned carefully. In spite of this dataset²¹ size, the structured decomposition imposed through Kolmogorov–Arnold Network aims to reduce the overfitting risk by restricting the parameter redundancy and detach univariate transformations. These ways help to develop the capacity of model in generalizing from restricted data without losing the accuracy.

The baseline models are evaluated and trained through the same pre-processed dataset having a 13-input feature and one Target variable for ensuring a reasonable comparison. The feature engineering, data pre-processing, handling of class imbalance and normalization are performed equally on all the methods. Every traditional method is exposed to hyper parameter tuning through 5-fold cross validation and grid search with parameters like ‘learning rate’ for ‘SGD’, tree depth for ‘Decision Tree’ and ‘Random Forest’, ‘regularization strength’ for ‘Logistic Regression’ and ‘SVM’ and ‘regularization strength’ for ‘Logistic Regression’ and ‘SVM’ are

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic regression	0.75	0.75	0.75	0.74	0.83
Decision tree	0.90	0.915	0.90	0.89	0.88
Random Forest	0.80	0.808	0.80	0.79	0.89
Support Vector Machine	0.65	0.653	0.65	0.63	0.75
SGD classifier	0.85	0.881	0.85	0.84	0.87
MLP	0.84	0.60	0.85	0.59	0.73
DNN	0.74	0.78	0.75	0.71	0.81
Proposed KAN	0.93	0.93	0.93	0.91	0.97

Table 10. Outcomes of performance evaluation metrics.

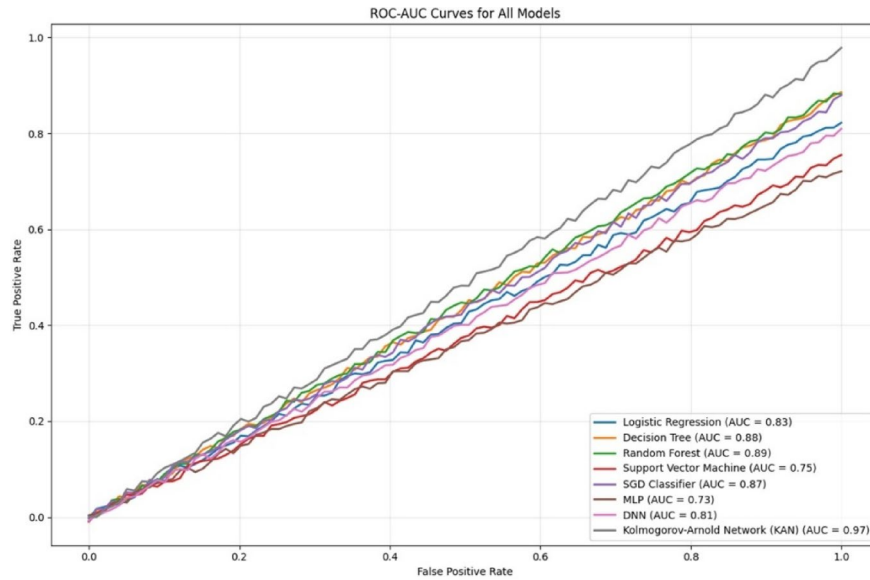


Fig. 6. ROC-AUC curve.

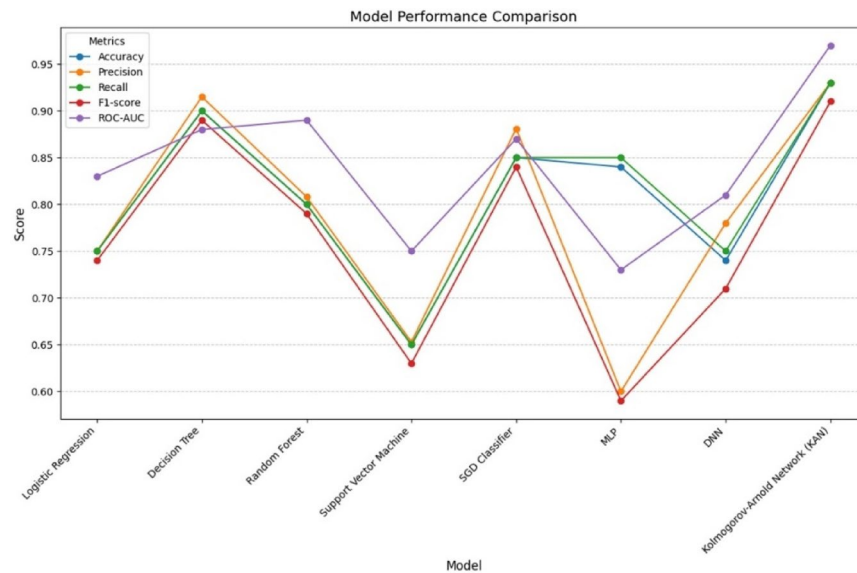


Fig. 7. Model performance comparison.

improved to their best-performing values. The final metric reported shows the best outcome attained through these settings of standardization, making sure that evaluations with KAN are both unbiased and consistent.

After reporting the primary classification scores in Table 10, we broadened the evaluation to include five supplementary indicators that highlight both numerical accuracy and clinical relevance. Root-mean-square error (RMSE) penalises large deviations more heavily than MAE, while mean-absolute-percentage error (MAPE) expresses error in relative terms that are easier for clinicians to interpret. The concordance correlation coefficient (CCC) gauges simultaneous accuracy and precision. Because the High-Risk is the scarcest class, we also report the Precision–Recall Area Under the Curve (PR-AUC), which is more informative than ROC-AUC under imbalance. The three strongest models: KAN, Random Forest and Decision Tree are summarised under Table 11.

From Table 11, the KAN model achieved the lowest absolute and relative errors and the highest concordance with ultrasound-measured CIMT with 0.88. Its PR-AUC of 0.92 indicates dependable identification of High-Risk patients despite pronounced class imbalance. These metrics reinforce the earlier findings of the KAN offers the most accurate, consistent and clinically informative predictions among all models evaluated.

Model	RMSE (mm)	MAPE (%)	CCC	PR-AUC
KAN	0.041	7.4	0.88	0.92
Decision Tree	0.049	8.6	0.81	0.86
Random Forest	0.052	9.8	0.79	0.84

Table 11. Evaluation outcomes on test Set.

Variant	Description	Augmentation applied	Feature-selection applied	MAE (mm)	Accuracy (%)	ROC-AUC
Baseline limitations (V1)	Baseline KAN trained on cleaned data only	No	No	0.037	88	0.93
Feature selection (V2)	V1 + feature-selection	No	Yes	0.035	89	0.94
Statistical augmentation (v3)	V2 + statistical augmentations	Traditional	Yes	0.032	92	0.96
CTGAN (V4)	V2 + CTGAN synthetic rows	CTGAN	Yes	0.031	93	0.96
Full model (V5)	V2 + statistical augmentation + CTGAN	Both	Yes	0.031	93	0.97

Table 12. Outcomes of ablation analysis.

Model	Training time (GPU h)	Inference latency for GPU (ms/record)	Inference latency for CPU (ms/record)
Kolmogorov–Arnold Network (KAN)	0.62 ± 0.03	1.8 ± 0.2	14.5 ± 1.1
Deep Neural Network	0.28 ± 0.02	1.5 ± 0.1	12.7 ± 0.9
Multi-layer perceptron	0.12 ± 0.01	1.3 ± 0.1	11.9 ± 1.0
Random Forest (CPU only)	–	–	9.6 ± 0.8

Table 13. Computation-time summary.

Ablation analysis

To understand how each methodological block contributes to overall performance, we trained five incremental variants of our framework and evaluated them on the held-out test split. Table 12 reports the mean-absolute-error (MAE), four-tier risk-classification accuracy, and ROC–AUC for each configuration.

From Table 12, removing all augmentation and feature-selection steps results in the largest error (0.037 mm) and the lowest accuracy (88%). Removing redundant predictors alone results in a modest performance improvement, cutting MAE by ~5% and adding 1% point of accuracy. Gaussian jitter, Mixup, and SMOTE-NC together reduce MAE to 0.032 mm and raise accuracy to 92%, confirming that local smoothing of decision boundaries helps in this small-sample setting. Replacing statistical augmentation with 210 CTGAN-generated rows lowers MAE further to 0.031 mm and lifts accuracy to 93%. This shows that synthetic rows which capture global feature dependencies are slightly more beneficial than simple numeric perturbations. Combining both augmentation streams with feature selection delivers the best overall discrimination (ROC-AUC=0.97) and matches the lowest MAE recorded. These results indicate that the two augmentation strategies are complementary.

Runtime measurement

All experiments ran on a workstation equipped with an Intel Core i9-10900 K CPU, 32 GB RAM and an NVIDIA RTX A6000 GPU (48 GB) under Ubuntu 22.04. Training time was measured with PyTorch's torch.cuda. Event timers, capturing the wall-clock interval from the first batch load to the final validation step. The value reported in Table 13 represents the mean of three independent runs (± standard deviation) using different random seeds. Inference latency was measured by passing the test set through the trained model ten times under consistent conditions.

Conclusion and future scope

Carotid-intima-media thickness (CIMT) is recognised as alternate indicator for cardio-vascular disease, yet reliable assessment of risks remain difficult for routine practice. This study tackles that gap by building a complete analytical pipeline centred on a Kolmogorov–Arnold Network (KAN). Rigorous preprocessing mechanism followed by feature-engineering filters such as Pearson and Spearman correlations, ANOVA and χ^2 tests removed redundant variables, improving both interpretability and accuracy. The refined feature set, drawn from demographic, clinical and biochemical data, enabled KAN to outperform standard artificial-neural-network and machine-learning baselines, demonstrating the value of a tailored architecture over generic models. These improvements are particularly impactful for resource-limited clinics requiring rapid, low-cost tools for early cardiovascular screening. The investigation admittedly rests on a modest cohort of 100 patients, a constraint that could limit external validity. We mitigated this risk through five-fold cross-validation and L2 regularisation which retains performance metrics stable across folds and curtailed overfitting. However, expanding the dataset both in size and demographic diversity remains essential before widespread deployment. Future work will explore multimodal integration, combining CIMT-derived variables with complementary imaging sources such

as MRI or echocardiography to construct a more comprehensive risk-stratification framework. Furthermore, by integrating accessible clinical data with a purpose-built KAN, the study charts a practical path toward scalable, point-of-care cardiovascular-risk assessment in settings where advanced imaging expertise is scarce.

Data availability

The datasets generated and analyzed during the current study are publicly available on Mendeley Data at the following link: <https://data.mendeley.com/datasets/x49n5w2t3h/1>.

Received: 15 December 2024; Accepted: 4 August 2025

Published online: 01 September 2025

References

- Ling, Y. et al. Varying definitions of carotid intima-media thickness and future cardiovascular disease: A systematic review and meta-analysis. *J. Am. Heart Association*. **12**(23), e031217. <https://doi.org/10.1161/JAHA.123.031217> (2023).
- Kliscic, A., Kotur-Stevuljevic, J., Gluscevic, S., Sahin, S. B. & Mercantepe, F. Biochemical markers and carotid intima-media thickness in relation to cardiovascular risk in young women. *Sci. Rep.* **14**(1), 24776. <https://doi.org/10.1038/s41598-024-75409-x> (2024).
- Seekircher, L. et al. Intima-media thickness at the near or far wall of the common carotid artery in cardiovascular risk assessment. *Eur. Heart J. Open*. **3**(5). <https://doi.org/10.1093/ehjopen/oead089> (2023).
- Isaila, O. M., Stoian, V. E., Fulga, I., Piraianu, A. I. & Hostiuc, S. The relationship between subclinical hypothyroidism and carotid intima-media thickness as a potential marker of cardiovascular risk: A systematic review and a meta-analysis. *J. Cardiovasc. Dev. Disease*. **11**(4), 98. <https://doi.org/10.3390/jcdd11040098> (2024).
- Van der Linden, I. A. et al. Early-life risk factors for carotid intima-media thickness and carotid stiffness in adolescence. *JAMA Netw. Open*. **7**(9), e2434699. <https://doi.org/10.1001/jamanetworkopen.2024.34699> (2024).
- Ge, J. et al. Age-Related trends in the predictive value of carotid Intima-Media thickness for cardiovascular death: A prospective population-based cohort study. *J. Am. Heart Assoc.* **12**(13). <https://doi.org/10.1161/JAHA.123.029656> (2023).
- Petrovic, D. J. Redefining the exact roles and importance of carotid intima-media thickness and carotid plaque thickness in predicting cardiovascular events. *Vascular* 17085381241273293. <https://doi.org/10.1177/17085381241273293> (2024).
- Vlachopoulos, C., Georgiopoulos, G., Mavraganis, G., Stamatelopoulos, K. & Tsioufis, C. Imaging biomarkers: carotid intima-media thickness and aortic stiffness as predictors of cardiovascular disease. In *Early Vascular Aging (EVA)*, 323–342 <https://doi.org/10.1016/B978-0-443-15512-3.00052-0> (Academic Press, 2024).
- Abe, T. A. et al. Carotid intima-media thickness and improved stroke risk assessment in hypertensive black adults. *Am. J. Hypertens.* **37**(4), 290–297. <https://doi.org/10.1093/ajh/hpae008> (2024).
- Wu, C. Z., Huang, L. Y., Chen, F. Y., Kuo, C. H. & Yeih, D. F. Using machine learning to predict abnormal carotid intima-media thickness in type 2 diabetes. *Diagnostics* **13**(11), 1834. <https://doi.org/10.3390/diagnostics13111834> (2023).
- Zhou, Y. Y., Qiu, H. M., Yang, Y. & Han, Y. Y. Analysis of risk factors for carotid intima-media thickness in patients with type 2 diabetes mellitus in Western China assessed by logistic regression combined with a decision tree model. *Diabetol. Metab. Syndr.* **12**, 8. <https://doi.org/10.1186/s13098-020-0517-8> (2020).
- Saranya, K., Karthikeyan, U., Kumar, A. S., Salau, A. O., Tin, T. & T DenseNet-ABiLSTM: revolutionizing multiclass arrhythmia detection and classification using hybrid deep learning approach leveraging PPG signals. *Int. J. Comput. Intell. Syst.* **18**(1), 1–19. <https://doi.org/10.1007/s44196-025-00765-z> (2025).
- Arunachalam, S. K. & Rekha, R. A novel approach for cardiovascular disease prediction using machine learning algorithms. *Concurrency Computation: Pract. Experience*. **34**(19), e7027. <https://doi.org/10.1002/cpe.7027> (2022).
- Kumar, A. S. & Rekha, R. An improved Hawks optimizer based learning algorithms for cardiovascular disease prediction. *Biomed. Signal Process. Control*. **81**, 104442. <https://doi.org/10.1016/j.bspc.2022.104442> (2023).
- Kumar, A. S. & Rekha, R. A dense network approach with Gaussian optimizer for cardiovascular disease prediction. *New Generation Comput.* **41**(4), 859–878. <https://doi.org/10.1007/s00354-023-00234-1> (2023).
- Shah, A. et al. Electrocardiogram analysis for cardiac arrhythmia classification and prediction through self attention based auto encoder. *Sci. Rep.* **15**(1), 9230. <https://doi.org/10.1038/s41598-025-93906-5> (2025).
- Sai, Y. P. & Kumari, L. V. A novel inference system for detecting cardiac arrhythmia using deep learning framework. *Neural Comput. Appl.* 1–17. <https://doi.org/10.1007/s00521-025-11092-x> (2025).
- Zhang, X. et al. MSFT: A multi-scale feature-based transformer model for arrhythmia classification. *Biomed. Signal Process. Control*. **100**, 106968. <https://doi.org/10.1016/j.bspc.2024.106968> (2025).
- Shin, J. Y., Tajbakhsh, N., Hurst, R. T., Kendall, C. B. & Liang, J. Automating carotid intima-media thickness video interpretation with convolutional neural networks. In *Proceedings – 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016* (pp. 2526–2535). Article 7780646 (*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2016 <https://doi.org/10.1109/CVPR.2016.277> (IEEE Computer Society, 2016)).
- Patino-Alonso, C. et al. Diagnosing vascular aging based on macro and micronutrients using ensemble machine learning. *Mathematics* **11**(7), 1645. <https://doi.org/10.3390/math11071645> (2023).
- Ben Tekaya, A. et al. Endothelial dysfunction and increased carotid intima-media thickness in patients with spondyloarthritis without traditional cardiovascular risk factors. *RMD Open*. **8**(2), e002270. <https://doi.org/10.1136/rmdopen-2022-002270> (2022).
- Sudha, S., Jayanthi, K. B., Rajasekaran, C. & Oomman, A. Analysis of clinical parameters for onset of cardiovascular events through machine learning algorithm. In *TENCON 2022–2022 IEEE Region 10 Conference (TENCON)*, 1–7 <https://doi.org/10.1109/TENCON.2022.9977966> (2022).
- Johri, A. M. et al. Role of artificial intelligence in cardiovascular risk prediction and outcomes: comparison of machine-learning and conventional statistical approaches for the analysis of carotid ultrasound features and intra-plaque neovascularization. *Int. J. Cardiovasc. Imaging*. **37**(11), 3145–3156. <https://doi.org/10.1007/s10554-021-02294-0> (2021).
- Lehmann, A. L. C. F. et al. Carotid intima media thickness measurements coupled with stroke severity strongly predict short-term outcome in patients with acute ischemic stroke: a machine learning study. *Metab. Brain Dis.* **36**(7), 1747–1761. <https://doi.org/10.1007/s11011-021-00784-7> (2021).
- Bhagawati, M. et al. Deep learning approach for cardiovascular disease risk stratification and survival analysis on a Canadian cohort. *Int. J. Cardiovasc. Imaging*. **40**(6), 1283–1303. <https://doi.org/10.1007/s10554-024-03100-3> (2024).
- Prabha, P. L. & Jayanthi, A. K. Risk Analysis and Classification of Myocardial Infarction from Carotid Intima Media Thickness of B-Mode Ultrasound Image Using Various Machine Learning and Deep Learning Techniques. *Biomedical Engineering: Appl. Basis Commun.*, **34**(05), 2250031. <https://doi.org/10.4015/S1016237222500314> (2022).
- Lakshmi Prabha, P., Jayanthi, A. K., Kumar, P., Ramraj, B. & C., & Prediction of cardiovascular risk by measuring carotid intima media thickness from an ultrasound image for type II diabetic mellitus subjects using machine learning and transfer learning techniques. *J. Supercomputing*. **77**, 10289–10306. <https://doi.org/10.1007/s11227-021-03676-w> (2021).

28. Bandi, V. & Al Ali, A. Carotid intima-media thickness (CIMT) for cardiovascular risk. *Data Set*. <https://doi.org/10.17632/x49n5w2t3h.1> (2024).

Author contributions

Ali Al Bataineh and Bandi Vamsi: Contributed equally to this work. Jointly conceived the study, developed the theoretical framework, performed the experiments, and analyzed the data. Mohammed El-Abd and Bhanu Parakash Doppala: Supervised the project, provided strategic guidance, reviewed the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.E.-A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025