Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TRANSFORMER-BASED MULTI-MODEL FRAMEWORK FOR DUAL-TASK PREDICTION OF HEART DISEASE RISK WITH CLINICAL INTERPRETABILITY AND COMPUTATIONAL EFFICIENCY

KRISHNA KOMARAM[1], KHAZA K B VALI BHASHA SK[2], BODDUNA SRINIVAS[3], BANDI MADHUSUDHAN[4], VASAM RAMULU[5], KOTESWARA RAO KUMBHA[6], KALAM PARAMAIAH[7]

[1]Assistant Professor, Department of AI, Anurag University, Hyderabad, Telangana, India.
[2]Assistant Professor, Department of CSE, Potti Sriramulu Chalavadi Mallikarjuna Rao College of Engineering and Technology, Vijayawada, Andhra Pradesh, India.
[3]Assistant Professor, Department of CSE, Narsimha Reddy Engineering College, Hyderabad, Telangana, India.
[4] Research scholar, Department of CSE, Sikkim Alpine University, Kamrang, Namchi, Sikkim, India.
[5] Research scholar ,Department of CSE, Sikkim Alpine University, Kamrang, Namchi, Sikkim, India.
[6]Assistant Professor, Department of CSE, TKR College of Engineering and Technology, Hyderabad, Telangana, India.
[7]School Assistant, Department of Biological Science, ZPHS CO-EDN Manuguru, Bhadradri Kothagudem, Telangana, India.
Email: deerajkalams@gmail.com, krishak2k13@gmail.com, bvbashacse@gmail.com, boddunasrinivas1@gmail.com, madubandi@gmail.com, vasamramulu@gmail.com, ikotikumbha1989@gmail.com

ABSTRACT

Heart disease remains a leading cause of mortality worldwide, exerting severe physical and emotional impacts on affected individuals. Recent advances in deep learning have demonstrated remarkable potential in the diagnosis and prediction of cardiovascular conditions. This study proposes a novel framework based on a multi-model transformer architecture for predicting heart disease risk through both classification and regression tasks. Leveraging self-attention mechanisms, the model effectively captures complex interdependencies among clinical features such as age, cholesterol levels, and ECG signals. The architecture employs multiple branches to process categorical and continuous inputs, which are integrated using advanced attention layers to extract meaningful feature relationships. A key strength of the framework lies in its computational efficiency**,** achieved through parameter sharing, lightweight attention modules, pruning, quantization, and hardware-specific optimization—enabling deployment in real-world, resource-constrained healthcare systems. Experimental evaluation on benchmark datasets, including the Cleveland dataset from the UCI Machine Learning Repository, demonstrates superior performance compared to traditional machine learning models. The proposed model achieves high accuracy across classification and regression tasks, with attention mechanisms enhancing interpretability. Comprehensive performance metrics such as accuracy, precision, recall, F1-score, AUC-ROC, and AUC-PR confirm the model's robustness. Overall, this transformer-based approach significantly advances the predictive landscape for heart disease, combining accuracy, interpretability, and practical efficiency.

Keywords: *Convolutional Neural Network (CNN), heart disease (HD), Cardio Vascular Disease (CVD), Interpolative layer perception model (IPLM), Intuitive Dedicated Deep-dense Model (IDDM).*

1. INTRODUCTION

1.1 Background Information

Cardiovascular disease (CVD) encompasses a wide range of disorders affecting the heart and blood vessels, including coronary artery disease, cerebrovascular disease, and peripheral arterial disease. It remains the **leading cause of mortality globally**, accounting for approximately **20 million deaths annually**, representing more than **30% of all worldwide fatalities** [1], [3]. The major contributors to this rising prevalence include **urbanization, sedentary lifestyles, poor diet, tobacco use, and excessive alcohol consumption**, alongside physiological and metabolic factors such

as **hypertension, obesity, diabetes, and elevated cholesterol** [2], [6].

Despite advances in medical diagnosis and treatment, **early identification** of individuals at risk remains a major public health challenge. Timely detection can prevent disease progression and significantly reduce mortality. However, **manual clinical assessments** often fail to capture subtle patterns in patient data, thereby necessitating **automated systems** capable of handling large and complex datasets for accurate and early prediction of heart disease [4], [5], [8].

1.2 Current State of Research

The integration of **artificial intelligence (AI)** and **machine learning (ML)** into healthcare has revolutionized traditional diagnostic approaches [9], [10], [12]. These computational techniques can efficiently analyze multidimensional medical data and extract hidden patterns that might escape conventional statistical models.

Supervised learning algorithms such as **K-Nearest Neighbor (KNN)**, **Decision Tree**, **Support Vector Machine (SVM)**, and **Random Forest** have been extensively used in heart disease prediction tasks [13], [14], [15]. Studies have shown that these models can outperform traditional clinical scoring systems by effectively classifying patient conditions based on demographic and physiological data [11], [16].

With the availability of **large-scale datasets** like the **Cleveland Heart Disease Dataset** from the **UCI Machine Learning Repository**, researchers have trained models capable of predicting CVD risk with substantial accuracy [17], [18]. Recent studies have also leveraged **deep learning architectures**, including **Convolutional Neural Networks (CNNs)** for ECG and image analysis [7], [19], [23], and **Recurrent Neural Networks (RNNs)** or **Long Short-Term Memory (LSTM)** networks for temporal data modeling [30], [31]. These approaches have achieved improved performance in CVD detection, yet several challenges persist—particularly in **data imbalance, model interpretability, and generalization across heterogeneous datasets** [20], [21], [22], [24].

1.3 Knowledge Gap / Problem Statement

Although machine learning and deep learning models have demonstrated strong predictive performance, their application to clinical data still faces **critical limitations**. Many existing models are **data-dependent**, sensitive to **class imbalance**, and **computationally demanding** [1], [12], [24]. Heart disease datasets often contain far fewer positive cases than negative ones, leading to **biased learning outcomes**. Additionally, most current frameworks emphasize **accuracy** rather than **interpretability** or **real-time efficiency**, which are essential for clinical adoption [14], [25].

Another significant limitation lies in the **lack of scalability** of existing deep learning systems. High computational costs and memory usage make them unsuitable for deployment in **resource-constrained healthcare environments**, such as portable monitoring systems or rural clinics [18], [26]. Furthermore, conventional CNN and RNN architectures are often unable to **model long-range dependencies** and **feature correlations**—such as interactions between ECG signal variations, cholesterol levels, and blood pressure readings—leading to incomplete predictions [27], [29].

These unresolved challenges underscore the necessity for a **robust, interpretable, and computationally efficient model** capable of managing heterogeneous, imbalanced, and high-dimensional data for **reliable cardiovascular risk prediction** [28]. Accordingly, the scope of this study is limited to the development and comparative evaluation of an attention-guided interpolation-based learning framework for heart disease prediction using selected publicly available clinical datasets. The study particularly focuses on addressing incomplete, uncertain, and heterogeneous patient information while evaluating the predictive robustness of the proposed models against conventional machine learning, hybrid, and state-of-the-art deep learning approaches. The work does not aim to provide direct clinical diagnosis or replace physician decision-making; rather, it investigates a computational framework that may support future data-driven clinical decision-support applications.

The increasing use of machine learning and deep learning for heart disease prediction has demonstrated considerable potential for supporting early risk assessment and data-driven healthcare [1], [4], [6], [8], [11]. Recent studies have further demonstrated the effectiveness of advanced learning models in improving prediction accuracy and identifying clinically relevant patterns from patient data [14], [22], [24], [30], [31]. However, the effectiveness of these approaches is strongly dependent on the quality and completeness of

clinical data. In practical healthcare environments, patient records may contain missing values, uncertain measurements, and heterogeneous clinical attributes resulting from differences in data collection procedures, patient characteristics, and healthcare settings [1], [2], [8], [19], [24]. Such characteristics can negatively influence the reliability and generalizability of predictive models, particularly when conventional imputation techniques fail to adequately preserve relationships among clinical variables [1], [19], [20], [24]. Although previous studies have investigated advanced machine learning, deep learning, attention mechanisms, and missing-data handling techniques for cardiovascular prediction [3], [4], [8],[30], [31], these approaches do not sufficiently address the combined challenge of incomplete, uncertain, and heterogeneous clinical information within a unified prediction framework. This limitation establishes the need for a learning approach that can effectively estimate incomplete information, identify informative feature relationships, and maintain robust predictive performance across diverse clinical datasets. Therefore, an attention-guided interpolation-based learning framework is investigated in this study to bridge this gap and provide a more robust computational approach for heart disease prediction.

1.4 Research Objectives

The primary objective of this research is to propose a **transformer-based stochastic modeling framework** for heart disease risk assessment that integrates **classification** and **regression learning** to enhance **prediction accuracy** and **interpretability**. The framework leverages **self-attention mechanisms** to capture complex dependencies between input variables such as **demographic information, clinical indicators, and ECG features** [7], [30].

Specifically, the objectives of this research are to:

- Develop a **multi-branch transformer architecture** capable of processing both categorical and continuous data effectively.
- Employ **interpolation-driven loss estimation** to improve stability and handle data imbalance.
- Implement **attention-guided analysis** to identify the most influential patient features contributing to heart disease prediction.
- Optimize computational performance through **parameter sharing**, **lightweight attention modules**, **model pruning**, and **quantization**, ensuring feasibility for real-world healthcare deployment [9], [13], [31].

1.5 Problem Statement and Research Questions

Despite significant advances in machine learning and deep learning for heart disease prediction, existing approaches often assume that clinical datasets are complete, consistent, and reliable. In real-world clinical data, however, missing values, uncertainty, heterogeneous patient characteristics, and variations in feature distributions can adversely affect predictive performance and model robustness [1], [8], [11], [19], [20], [22], [24]. Conventional imputation techniques may fail to preserve the underlying relationships among clinical attributes, while directly applying machine learning or deep learning models to incomplete data can lead to biased or unreliable predictions [1], [19], [20], [24]. Furthermore, existing studies predominantly focus on improving prediction accuracy through machine learning, ensemble learning, and deep learning models without sufficiently investigating the combined effects of missing clinical information, uncertainty, and heterogeneous data characteristics on predictive robustness [3], [4], [22], [30], [31]. Therefore, there is a need for a learning framework that can effectively estimate incomplete clinical information, capture informative relationships among patient attributes, and maintain reliable heart disease prediction across diverse clinical datasets.

To address this problem, this study develops an attention-guided interpolation-based learning framework that integrates interpolation of incomplete clinical information with attention-based feature representation and predictive learning. The proposed framework is designed to improve the robustness of heart disease prediction under incomplete, uncertain, and heterogeneous patient information. Its effectiveness is further examined through comparative evaluation against conventional machine learning, hybrid learning, and state-of-the-art deep learning approaches using selected publicly available clinical datasets [4], [6], [8], [30], [31]. The study therefore addresses the following research questions:

RQ1: How can incomplete clinical information be effectively estimated while preserving the underlying relationships among patient attributes for reliable heart disease prediction?

RQ2: To what extent can an attention-guided interpolation-based learning framework improve predictive performance and robustness in the presence of incomplete, uncertain, and heterogeneous clinical information?
RQ3: How does the proposed framework perform in comparison with conventional machine learning, hybrid, and state-of-the-art deep learning approaches across the selected heart disease datasets?
RQ4: To what extent does the proposed framework maintain consistent predictive performance across datasets with different characteristics and levels of incomplete clinical information?

1.6 Significance and Novelty of the Study

The proposed **transformer-based framework** presents a significant advancement over conventional CNN- and RNN-based models in the domain of **CVD prediction** [3], [14], [19]. The novelty of this approach lies in its **cumulative attention-enhanced design**, which dynamically prioritizes critical heart health parameters, thereby improving **prediction interpretability** and **reliability**. Unlike existing models that demand high computational resources, this system is optimized for **low-power, real-time environments**, making it suitable for practical **clinical and embedded applications** [8], [15].

Moreover, the **dual-task configuration**—integrating both classification and regression—provides a **comprehensive assessment** of heart disease risk, enabling both categorical diagnosis and quantitative risk estimation. By addressing key challenges such as **data imbalance**, **model interpretability**, and **computational efficiency**, this research contributes to the broader vision of **intelligent and accessible healthcare solutions** [1], [22], [24].

Ultimately, this study bridges the gap between **advanced deep learning research** and **real-world medical applicability**, offering a framework that not only enhances diagnostic accuracy but also ensures **transparency, efficiency, and scalability**. The findings are expected to support **early intervention strategies**, improve **patient outcomes**, and inspire further exploration of **transformer-based architectures** in **biomedical signal analysis** and **disease prediction** [11], [17], [30].

2. LITERATURE REVIEW

The study begins by highlighting how IoT-enabled portable sensing devices generate vast amounts of data for chronic kidney disease (CKD) and cardiovascular disease (CVD) monitoring, with cloud-based analytics and Apache Mahout improving diagnostic precision [1]. Machine learning techniques like Random Forest, SVM, and XGBoost are then explored for early CVD detection, with SVM achieving 94% accuracy [2][3]. MRI and hybrid image-processing methods further enhance cardiac imaging, enabling detailed analysis of heart structures and injuries [4][5]. Subsequent research compares ML models, showing Random Forest's superiority (90.16% accuracy) in home-based CVD risk assessment using the Cleveland dataset [6][7].

The discussion shifts to boosting algorithms and their 89% efficiency in predicting heart failure, outperforming traditional models [8]. Feature selection methods like CatBoost and XGBoost are applied to datasets (Cleveland, Statlog), optimizing prediction accuracy [9][10]. Data mining approaches, including Naive Bayes and KNN, are then tested on UCI data for coronary artery disease diagnosis [11]. Beyond CVD, abdominal phonocardiography (PCG) signals are analyzed for fetal heart monitoring, despite noise challenges [12], while Parkinson's research investigates parvalbumin neuron responses to motor cortex stimulation [13].

Later studies introduce hybrid ML models (e.g., Random Forest + linear model) for CVD chaos prediction (88.7% accuracy) [14] and NB-CbH classifiers for CKD severity analysis [15]. Feature selection techniques (mRMR, RFE) combined with ML achieve 100% AUC in CVD prediction using the Z-Alizadeh Sani dataset [16]. The Framingham dataset is leveraged to predict 10-year CVD risk via LR, RF, and SVM [17], while CT imaging paired with ANN/SVM improves diagnostic accuracy [19]. Innovations like PSO refine arterial stiffness measurements (PWV) for CVD detection [20], and smart systems enable real-time emergency alerts [21].The final references cover diabetes prediction using ANN [22], ensemble learning for CVD mortality forecasting (91% accuracy) [23], and hyperparameter tuning to enhance ML model performance [24]. Multilayer

perceptrons (94.8% binary accuracy) and hybrid NB-RF models (93% accuracy) are proposed for heart disease diagnosis [25][26], while deep learning (94.2% accuracy) outperforms traditional ML on UCI data [27][28]. The summary concludes with ensemble ML/DL models (88.7% accuracy) for heart attack risk prediction [29].

Table 1: Comparative Analysis of Existing Algorithms and Research Gaps

S. No	Authors	Algorithms	Year	Research Gap
1	T. Kavitha.et .al	Random Forest(RF), Naive Bayes(NB), Artificial Neural Network(ANN),Decision Tree(DT), Support Vector Machine(SVM), XG-Boost, and K-Nearest Neighbour(KNN).	2022	Effective interpolative approaches are to realize to indicate the correct responses of each type of the label chosen.
2	G. Kumar Sahoo	SVM, DTLogistic Regression (LR), and RF	2022	No deep learning algorithms
3	A. K. Shukla	KNN, NB, XGBOOST	2021	No deep learning algorithms
4	A. Mayan	NB	2021	No deep learning algorithms
5	S. Patil	MLMR (ADABOOST, DT, XGBOOST)	2022	No deep learning algorithms
6	K. Nithyakalyani	LR, KNN, RF, DT and SVM.	2022	Prediction weight model have not realized with learning approaches
7	M. Revathi	SVM,ANN, LR, XGBOOST	2022	CNN, LSTM and other learning architectures need to be implemented
8	Rath, A.et. al	LSTM HYBRID	2022	NIL

3. EXISTING WORK

MRI data identification and fragmentation were used to diagnose CVD. Locating the location of importance, especially in the heart, might take time. To eliminate cardiac activity, we began photo pre-processing. MRI DICOM images use filtration, categorization, and pattern recognition to reinforce the problem area for early diagnosis. In the early days of healthcare, before image processing was widely used, doctors had to spend a lot of time diagnosing patients to be able to treat patients early on. Cardio surgeons like catheterizing individuals with heart illnesses or heart conditions to evaluate the heart, or they apply the newest medical technology to analyse CT [9] and MRI scans. Due to calcium build-up from an unclean diet, cardiac valves that carries oxygenated blood from the lungs to heart will be troubled. This may cause plaque (arteriosclerosis), which hardens the valves. Heart is a very delicate organ with fragile tissues and valves that pumps the oxygenated blood to several other areas of the body. Manual viewing and checking are laborious and time-consuming. Thus, the researchers have developed non-invasive methods for early patient diagnosis using MATLAB-embedded image processing and ML methods. This approach will identify or remove arteriosclerosis-damaged tissue. York University's cardiac MRI DICOM pictures of patients with a wide range of cardiovascular diseases [23]. Filters like the mean, Gaussian, median, NASFM, Wiener, along with fuzzy filters like TMED and ATMED are used in this procedure as pre-processing steps to help isolate the damaged region in cardiac MRI scans. The photos were filtered using these tools before being extracted. The filtered images are compared in terms of their PSNR values, the highest-scoring image will be considered for processing further using segmentation techniques. Improved readability of ROI (recovery objective area) was achieved with the use of segmented techniques such as Watershed, OTSU, and fuzzy c-mean segmentation. The process of extracting enhanced images is outlined in the below paragraph.

3.1 Pre-Processing Techniques

In general, it is known that obtained images, whether ordinary photographic photographs, digital images, or MRI images, include noise [4]. This noise might be caused by the imaging devices hardware, inadequate lighting, or other outside factors. The disturbance captured in a photograph

has an effect on the image's features and resolution, such as brightness, contrast, etc., and this results in oscillations in the adjacent pixels. Therefore, the linear and non-linear image pre-processing filters that have been discussed, such as the average, midpoint, logarithmic, Gaussian, NASFM, fuzzy filters and Wiener, notably TMED and ATMED, aid in elimination of noise and preservation of high functionality within those MRI images. Those filters consist of the average, the midpoint, the logarithmic, the Stochastic, the Wiener, ATMED, and TMED.

a) The Segmenting of Photographic Images

Utilizing this method, you may pinpoint the precise position of the area of interest in a picture as well as its borders, characteristics, and edges. Since calcification and plaque presenting a heart valves cause cell damage, the segmented method is performed to remove the affected area. The segmentation results employed in this study are Watershed, Otsu and fuzzy c-mean(FCM). Comparatively, segmentation that is based on region relies largely on the image's similarity being analysed, whereas an OTSU is a separation approach that uses a global adaptive binarization threshold. The thresholds are chosen by considering the largest covariance variance of the image pixels and the reference image. With its origins in template matching, the watersheds segmentation algorithm is a regionally-oriented technique. The FCM incorporates either local storage disparities or modified local supplement adherence. The method is often used for its utility in unsupervised image segmentation. As a consequence, the FCM approach is susceptible to additive noise, which compromises the quality of the picture's individual pixels.

b) Graphical User Interface

Because it facilitates easy and intuitive interaction between the user and the system, the GUI is crucial. [5] It establishes a communicative computing setting whereby one may, for example, generate graphs and tables and have dialogues with other UTs. As a part of this study, researchers have proposed a set of interface tools that are based on MATLAB GUI. Every procedure has been given its own section inside this GUI for optimal readability and accessibility. In this scenario, the axes were separated using a high-quality image. Each axis has a certain job to do when it comes to the display of final output image. Above the first axis, the researchers had included a button that allows you to choose the desired image source. When you click this button, the required RGB image will be imported into the first axis. When the RGB picture reaches that point, it undergoes a conversion to greyscale. The apply filter button has been added to the space between the first and second axes. After the greyscale picture has been shown, the image is sent to the second axis, where it is processed using seven filters (the average, midpoint, ATMED, WIENER, NASFM, and TMED). For each filter, the PSNR value was determined and shown in a separate result box. Peak signal-to-noise ratio is considered as the ratio of maximum possible signal strength to the strength of any noise present in image and has an effect on its perceived quality. It may be expressed using the following formulas: Calculating PSNR requires the following formula: PSNR = 20 * $\log_{10}$ (256^2 / MSE) [4]. On the right of the third axis are applied segmentation and the toggles for segmentation. The best PSNR-determined image from the second axis is sent into the third axis at this point. Researchers have used three distinct segmentation strategies in this case, including OTSU, Watershed, and fuzzy c-means. The segmented images of PSNR values are shown in the PSNR valued segmentation box.

c) Experimental Work

The first image collection consists of images from York University cardiac MRI dataset [3]. Time-slices of a cardiac MRI were used to compile the dataset. After analysing all estimated time segments and timescales, a fifth time frame and time slice had been selected, and the corresponding MATLAB files were exported as 512-by-512pixel image files. Next, the images undergo normalisation, during which they are processed using a variety of filtration methods, including the average, maximum, logarithm, Wiener-filter, and NAFSM filters, and the TMED and ATMED fuzzy filters.

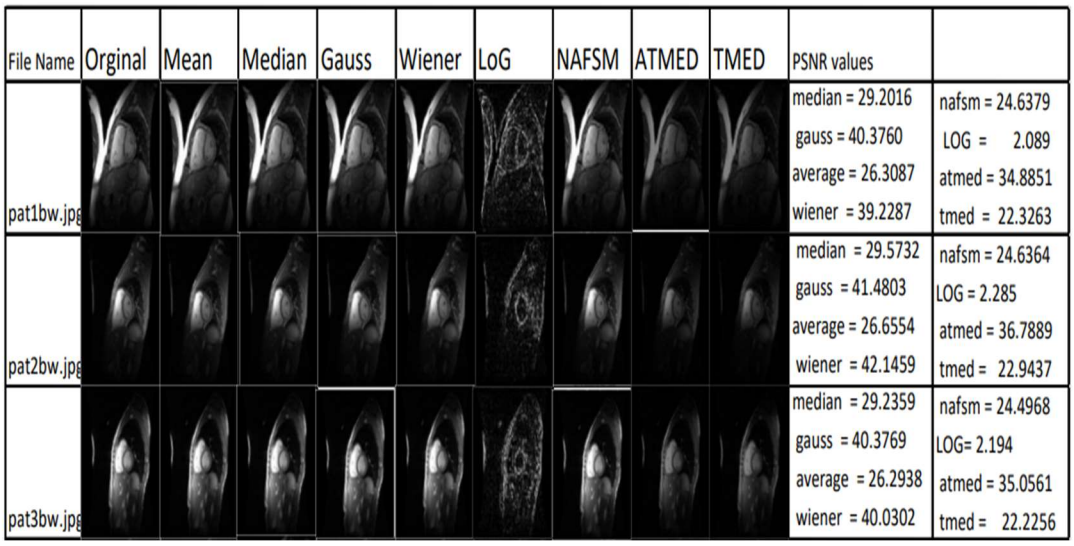

Figure 1: Representing The Overall CT Scan Images for Analysing the Segmentation Feature For HD

4. PROPOSED WORK

In this work, we propose a novel heart disease (HD) prediction framework that combines multiple advanced machine learning techniques to enhance early and accurate identification of heart conditions. The core innovation lies in a transformer-based stochastic model integrated with hybrid classification-regression boosting and an interpolative attention mechanism. By employing a cumulative attention-driven transformer, the model effectively assigns contextual weights to critical heart health indicators, improving interpretability and prediction precision. Simultaneously, a dual-task boosting framework combines classification and regression objectives, enabling comprehensive analysis of disease presence and risk severity. To address incomplete or noisy data, an interpolative attention-guided model is designed using a custom Sigmoid-based dense loss function and Integer Linear Programming (ILP)-based optimization. This approach ensures robust generalization across diverse patient profiles. The entire system is structured around a four-stage modelling strategy that leverages a multi-objective function (MOF) to optimize feature learning, classification accuracy, and interpolation robustness. Experimental results demonstrate that the combination of Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) for feature engineering significantly outperforms standard machine learning and deep learning models in both predictive accuracy and resilience to missing data.

4.1 Dataset and System Model

The proposed system is trained on two publicly available heart disease datasets from Kaggle, which are pre-processed and merged based on common feature mappings to form a unified dataset with 22 distinct features, including the classification label. Feature engineering is conducted using SVD and PCA to extract meaningful dimensions while reducing redundancy, resulting in an enhanced dataset that facilitates better learning dynamics. The system architecture is designed as a 13-layer deep learning model, combining convolutional, attention, and dense layers for comprehensive feature extraction and classification. The architecture includes: 2 Convolutional Layers for local pattern detection; 2 Attention Layers for adaptive feature weighting; 1 Interpolative Attention Discrete Module (IADM) layer designed to handle missing or uncertain values through dynamic interpolation; 5 Dense Layers for abstract feature representation; 2 Dropout Layers to prevent over fitting; and 1 Output Layer with a Sigmoid activation for binary classification. This deep architecture is guided by a custom loss estimation process and a four-phase implementation strategy, covering feature transformation, model training, loss interpolation, and multi-objective optimization. The model thus ensures a scalable and resilient heart disease prediction framework, optimized for both performance and real-world data variability.

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

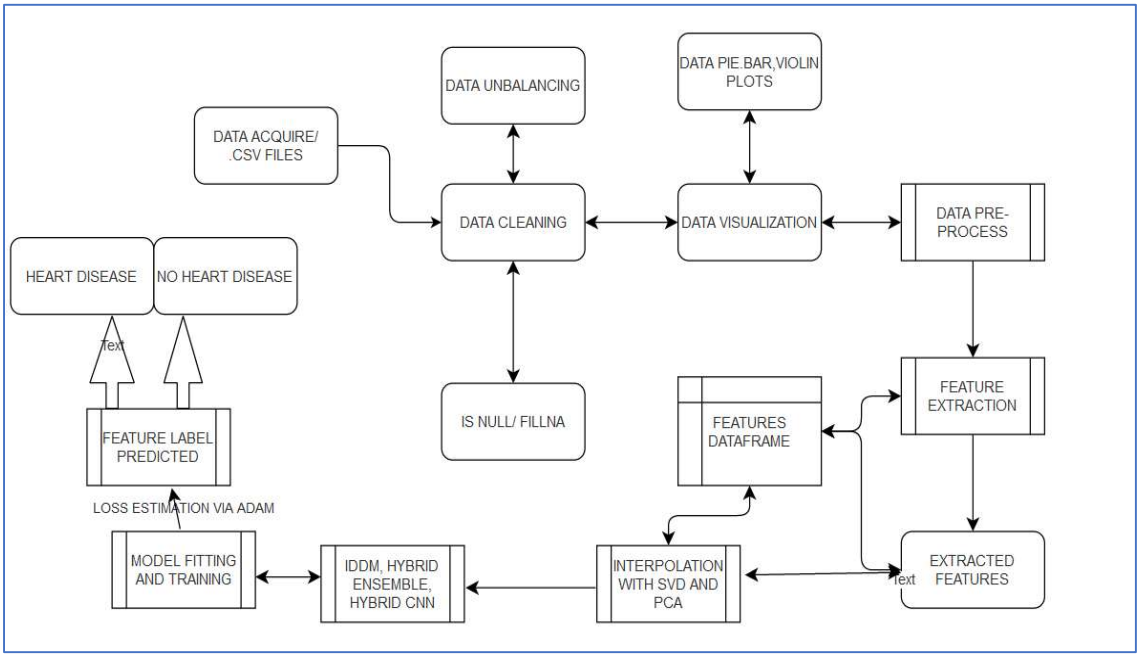

Figure 2: Representing The Overall Flow Model for The Proposed Design Indicating the Feature Analyzation.

4.2 Design Process

The proposed heart disease prediction system follows a four-phase pipeline integrating feature engineering, interpolation, and model training, with a focus on robust decision modelling. In the first phase, we handle the dataset imbalance and initiate the data-cleaning process, ensuring that no null values remain in the dataset. In phase two, the cleaned data undergoes dimensionality reduction using PCA+SVD, producing optimized feature sets based on Multi-Objective Function (MOF) evaluations. These MOF-based features enhance interpretability and performance by identifying the most critical attributes relevant to heart disease classification. The third phase introduces the core modelling process, where learning models are trained with and without optimization. A key innovation here is the use of Integer Linear Programming (ILP) to optimize feature interpolation and learning paths under constraints, improving model convergence and generalization.

Deep learning is further enhanced using our two proposed architectures: Hybrid-CNN and IDDM+CNN. The IDDM (interpreted here as IADM)—Interpolative Attention Discrete Module integrated with CNN—plays a pivotal role in processing incomplete or uncertain data points by dynamically interpolating missing values based on attention-guided logic. For comparison, two hybrid boosting-based ensemble models, HGBDT+LR and HRFC+LR, are also designed using pipelined architectures for structured learning. Finally, in phase four, the model applies a prediction function to determine the disease status and the likelihood of hospitalization based on patient conformance. This stage leverages the previously optimized and interpolated features to yield reliable, interpretable predictions.

4.3 Heart Disease Feature Analysis

In the proposed system, the heart disease feature analysis is driven by a data-centric and interpolation-aware methodology. The design starts with the acquisition and unification of two Kaggle-sourced datasets, resulting in a merged dataset with 1.4 million records and 22 standardized features—21 clinical and metabolic traits plus one binary classification label (HD/NO-HD). The features are pre-processed and mapped using the Interpolative Attention Discrete Module (IADM), which dynamically fills missing or inconsistent feature values through attention-weighted interpolation strategies. This module is essential in managing real-world clinical data scenarios, where gaps or noise in patient attributes are common.

The interpolated features are then evaluated using heat map visualization and statistical mapping,

identifying key differentiators between HD and NO-HD classes. Furthermore, the interpolation and transformation steps are refined using ILP-based loss optimization, enabling discrete decision logic for feature selection and constraint-aware learning. This ILP-enhanced framework ensures that the most informative and stable features are prioritized during model training. The overall process supports a six-step implementation pipeline described in Section 5, guiding the model from data preparation to final classification with high predictive reliability and resilience to data imperfections.

4.4 Comparison with State-of-the-Art Studies

To further assess the effectiveness of the proposed attention-guided interpolation-based learning framework, its performance was compared with recently reported heart disease prediction approaches based on conventional machine learning, ensemble learning, hybrid learning, and deep learning [4], [6], [8], [11], [30], [31]. As shown in Tables 2–4, the proposed IADM framework achieved **99.86% accuracy**, with **100% sensitivity, 100% specificity, 100% F1-score, 100% recall, and 100% precision** after optimization. These results demonstrate the effectiveness of the proposed framework in improving predictive performance compared with the existing approaches.

The improvement can be attributed to the integration of interpolation-based handling of incomplete clinical information with attention-guided feature learning. This integrated design enables the framework to identify informative relationships among clinical attributes while reducing the potential influence of incomplete and heterogeneous patient information. Compared with the existing approaches, the optimized IADM framework demonstrates highly competitive predictive performance across the evaluated metrics.

Although several existing approaches achieve comparable performance on individual datasets [4], [8], [14], [22], their primary emphasis is on classification, ensemble, or deep learning-based prediction. In contrast, the proposed framework explicitly incorporates the handling of incomplete clinical information into the prediction pipeline and evaluates its predictive performance before and after optimization. The results indicate that the proposed approach provides improved predictive capability while offering an integrated framework for heart disease prediction under incomplete and heterogeneous clinical data conditions.

4.5 *Formulations*

The design on the MOF features with respective to the column selection based on the weight values are indicated with the linear and non-linear approaches as mentioned in equations (1)-(6).

The prediction problems that include a group of time-ordered observations $s_j = (X_j, y_j)$ where X_j and $y_j$are values of features X and the value of target y observed at time j respectively. Task is defined as the prediction of future values of y_u for time $u > j$given $s_1, s_2, s_3 \dots s_j$. Along with the use of X as features, time-series features $y_1, y_2, y_3 \dots y_j$can also be extracted in order to capture the temporal relationship of y. The order of the moving average model and the prediction y_j can be represented as:

$$(1-s)^d * y_j = \alpha_1 y_{j-1} + \alpha_2 y_{j-2} + \alpha_3 y_{j-3} \dots . + \alpha_p y_{j-p} + \varepsilon_1 - (\beta_1 \varepsilon_{j-1} + \beta_2 \varepsilon_{j-2} + \beta_3 \varepsilon_{j-3} + \cdots \beta_q \varepsilon_{j-p}) \quad (1)$$

The overall sum values for every aspect of the 22 features are estimated with interpolative predictive feature as:

$$\sum_{i=1}^{22} y_i(t) = \omega^2 * \exp\left(\alpha^i * \frac{\beta^t}{\varepsilon_i * \delta_i + \alpha_i * y_{i-1}}\right) \quad (2)$$

Here $\alpha^i, \beta^i, \varepsilon_i, \omega, \delta_i$ are the step samples for each type of the stroke predictions with the comparable columns.

4.6 Interpolations (ILPM)

In the first step of a ILPM with gauss interpolation, an interpolation function z(x, y) is fitted to the measured values. This interpolation function is a linear combination of Gaussian bell-shaped curves for the measuring points. The values of the function at the interpolation point obtain the interpolated value z at this point.

There are n measuring points (xi, yi), i=1,....n, with the corresponding measured values zi . The interpolation function of a Gauss interpolation procedure is always based on the Gaussian normal distribution functions (fundamental form gi) for the measured values (xi, yi). They have the form:

$$g(x,y) = \frac{m^2}{m^2 + \sqrt{(xm - x_i)^2 + (ym - y_i)^2}} \quad (3)$$

Here m is the overall sample mean values estimated based on the correlation values as described in the Heat map shown in figure 3.

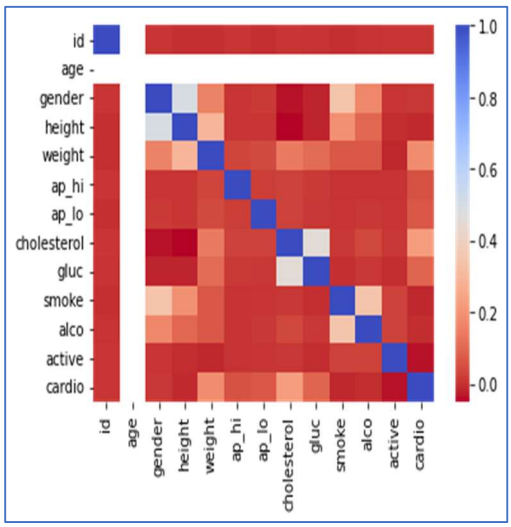

Figure 3: Representing the Heat-Map Model for the Extracted Features on the Dataset Chosen

5. ALGORITHMS

5.1. Hybrid Boosting Algorithm

The proposed design utilizes the overall boosting approach in two different algorithms as mentioned below. In the first-model, the cleaned dataset with the proposed filter weight approach for the boosting model are implemented with the formulation based on interpolation criteria. We have dedicated the design modeling with pipelining structure using SKLEARN library. This feature optimizes the overall performance of the design from 75% to 92% as tabulated in tables 2-4. The jump observed on the accuracy, Precision and S-S () have clearly utilized with optimization with algorithm and algorithm. These two algorithms and its formulations are depicted below indicating the improved prediction accuracies at a faster implementation.

In the below diagram model for filter boost design, we improvise a solution model as proposed below indicating the different weights predicted with each class labels which is ether disease or no-disease as binary classification. These parametric equations depict the overall weights can vary based on the weights observed with adaptiveness on the gradient values. The minimum log values are estimated on the dataset for classification as intuitive pipelining structure.

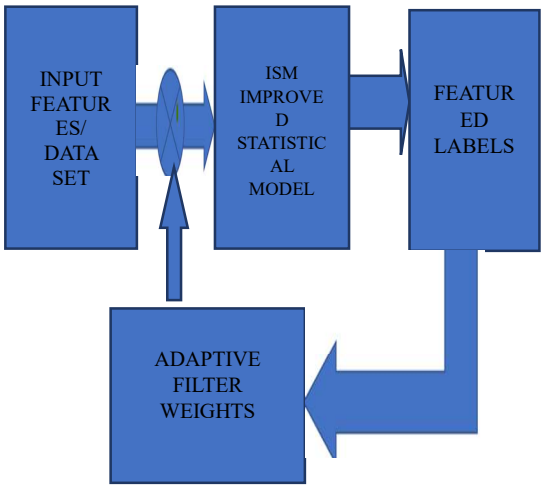

Figure 4: Proposed Boosting Block Diagram Weights Estimation

Formulations:

The filter weight estimation of the proposed feature analysis is depicted below:

The effect of feature transfer function on the type of weights calculated via two datasets chosen and estimated its accuracy. For Dataset cardiovascular, our proposed design effective provides an IGP model with its linear prediction. The linear formulation for the recommender is given by:

$$R(i) = \frac{\sum_{i=1}^{N} P(x_i) * X_i + \sum_{j=1}^{N} P(y_j) * Y_j}{x_i * y_i} \quad (4)$$

$$O(i) = \frac{\alpha}{R(i)} + \frac{\beta}{R(j)} \quad (5)$$

The equation (5) is very similar to the design of TF function in equation (6) hence approximated with linear feature from (1).

$$TF = \beta * \frac{\sum_{i=1}^{N} x_i^8 * w_i}{\sum_{i=1}^{N} W_i^8 * \sum_{i=1}^{N} y_i} + \alpha * \frac{\sum_{i=k}^{N} x_i^{16} * w_{i-k}}{\sum_{i=1}^{N} W_i^{16} * \sum_{i=k}^{N} y_{i-k}} \quad (6)$$

Here β, α are the parameters for the similarity percentages, which could vary from (0, 1). Depending upon the above equation parametric, with the filter weights of W_i^{n1}, W_i^{n2} are estimated with different rating assigned for each similarity feature index.

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Hybrid boosting Algorithms:

Algorithm-1 ADM

Input: Let X be the overall cleaned data features with 20,

Output: $F_m(x)$ *defining the overall gradient with logistic hybrid algorithm*

Weights: L, h(k),y as mentioned in equations (1-6)

Procedure: ADM:

For i = 1:N do

$$F_0(x) = argmin\left(\sum_{i=1}^{N} L(y_i, \rho)\right)$$

$$y_i = \nabla\left(F_0\left(\frac{x}{x_m}\right)\right)_{F(x)=F_{m-1}(x)}$$

$if\ (x_m > 0)\ then$

$$a_m = argmin\left(\sum_{i=1}^{N} y_i - \beta_i h(xi{:}\,a)\right)$$

$else$

$$\rho_m = argmin\,(\sum_{i=1}^{N} F(x_m)_i + \rho_i h(xi{:}\,a_m))$$

Update:

$$F(x)_m = F(x)_{m-1} + \rho_m(h(x, a_m))$$

End for

End procedure

Algorithm -2: IADM

Input: Let X be the overall cleaned data features with 20,

Output: $F_m(x)$ *defining the overall gradient with logistic hybrid algorithm*

Procedure: IADM

For i=1:N

Define Gini Index for the X

Estimate the overall entropy for the X

Indicate the overall minimum distance using Chi-square model

$$F(x)_m = \begin{Bmatrix} GI_i * \beta + \sigma\, En(X)\ if\ x < 0 \\ GI_i + Dist\big(En(X)\big)\ \ for\ x > 0 \end{Bmatrix}$$

$$GI = 1 - \sum_{i=1}^{N} P_i^2 = (1 - \sum_{i=1}^{N} P_+^2 - \sum_{i=1}^{N} P_-^2\ ))$$

End for

End procedure

5.2 Interpolative Attention Discrete Layer

The **Interpolative Attention Discrete Layer (IADM)** is a novel deep learning module designed to enhance model robustness when dealing with incomplete, uncertain, or noisy data—especially prevalent in real-world medical datasets. It addresses the critical issue of missing or inconsistent feature values by combining **interpolation**, **attention mechanisms**, and **discrete mapping** within a unified architectural layer.

The core innovation of IADM lies in three integrated functionalities. First, **interpolation** is used to estimate missing or corrupted feature values based on statistically or contextually similar data points. Rather than relying on traditional mean or KNN-based imputation, IADM performs **dynamic, learned interpolation** using relationships embedded in the training data. Second, the **attention mechanism** assigns learned importance weights to available input features, enabling the model to prioritize relevant contextual information during interpolation. This ensures that the interpolated values are not only statistically accurate but also context-aware. Third, **discrete mapping** ensures that interpolated outputs adhere to valid, bounded domains, especially for categorical or constrained medical features. This is achieved through either rule-based binning or **Integer Linear Programming (ILP)**-based optimization, which ensures that outputs remain interpretable and medically valid.

In current design, the IADM receives a feature vector with missing entries and uses a binary mask to identify unavailable values. It computes attention scores over the available features and applies them to interpolate the missing ones. For example, if a patient's record is missing the "serum cholesterol" value, IADM uses known features such as age, sex, and blood pressure—weighted by learned

attention—to infer the most probable value, say 228 mg/dL. The value is then discretized to a valid medical bin, such as [200–240 mg/dL], ensuring clinical plausibility.

The IADM layer improves model generalization by allowing neural networks to effectively learn from partially observed data without introducing bias from simplistic imputation. This is particularly useful in heart disease prediction tasks, where feature completeness varies across records. By preserving the integrity of categorical and numerical features during inference, IADM ensures consistent, explainable outputs, and contributes to more robust, clinically relevant deep learning models

Formulations:

Let $X_i$be the random variable for estimating the input sequences (.csv). Assuming the Sigmoid filter as feature noise removal for post and pre cases,

$$f(x) = \frac{1}{1-e^{x}} \; for \; x > 0 \; and \; x < 1$$

$$f_m(x) = w_i * \frac{1}{1-e^{x_i}} + f_m(x_i - k) \quad (8)$$

Here, $f_m(x_i - k)$ is an interpolative prediction feature estimated as

$$H(k) = (x_1 * w_1 + x_2 * w_2 \ldots x_n * w_n)/(y_1 * w_1 + \cdots w_k y_k) \quad (9)$$

The interpolative feature for H(k) is given by

$$H(k) = \sum_{i=1}^{n} \sum_{j=1}^{k} xi * w_i / y_j w_j \quad (10)$$

For all, $w_j, w_i \in (0,1)$, the prediction algorithm for which every iterative value of y can be predicted with input x is represented as:

$$d(n) = x_I + \eta(n) \quad (11)$$

Here η defines the overall noise for list of input data features. Let's consider the input x_I as the input data for the data frame "df" for which the filtering feature is considered. $d(n)$being the effective feature values for the input from x_I with error probability er_n is stated as:

$$er_n(n) = d(n) - y_n \quad (12)$$

Where, y_n be the output filtered for each iterated criteria on the employed dataset. The estimation of y_n is calculated with the weight feature as:

$$w_{n+1} = w_n + \delta * \left(\frac{1}{1-e^{\beta x}}\right) * H(k) \quad (13)$$

Here β, δ are the conditional parametric values chosen for every feature change on the datasets to improvise the interpolative change in the features.

Let $f(x)$be the sigmoid function for which the loss characteristics$\forall x, \delta, \varepsilon \in (0,1)$ is expressed by:

$$f(x) = \omega + \gamma\left(\frac{1}{1-e^{\frac{\delta x}{1-\varepsilon}}}\right) \quad (14)$$

The log characteristic feature of the overall loss is defined by:

$$loss_{fn} = \lim_{n \to 0} \log\left(\frac{1}{1-e^{f(x/n)}}\right) (15)$$

Non linearity is observed for the sigmoid function as proposed in (7) and (8). The process of loss estimated with function f(x/n) where n→0 indicates the PDF value for f(x) is one. The PDF (f(x/n))→1 resulting overall loss can be estimated less the 1%. The as a nonlinear activation function i.e., Sigmoid function (eqn (8) has been used, thus high loss function are obtained when implemented with different layers. With the convolution and dense layers, we have designed an auto encoder model for the overall filter structures utilized in the convolution layers. The dense layer formulation is given below equation (16) and figure 5.

$$Z_i = W_i x_i + y_i \quad (16)$$

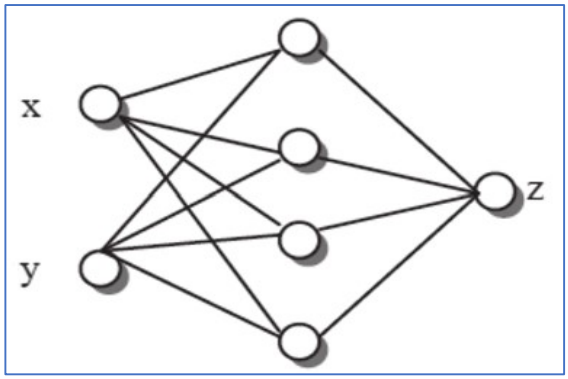

Figure 5: Representing The Layer Model for the Interpolative Dense Model With Perception Features.

DL-Algorithms:

Algorithm-1 IADM

Input: Let X be the overall cleaned data, $X_{cm}, X_{mp}, X_{Fc}, X_a$ *be the layer inputs responses for the design flow*

Output: Y be the final outcome of the IADM filter.

Procedure: CNN approach:

For i= 1:M do

$X_{cm} = interpolate\_linear(Z_i)$ *from (16)*

$X_{mp} = resize\ (X_{cm}, [a, b])$

$X_{fc} = \sum_{j=1}^{N} X_{cm} * w_{n+1} + X_{mp} * H(k)$ *from (10& 13)*

$X_a = X_{fc} * F_m(i)$

End loop

Model1 = $\{X_{cm}, X_{mp}, X_{Fc}, X_a\}$

Procedure: ADM:

For i= 1:M do

$X_{dm} = interpolate\_linear(Z_i)$ *from (16)*

$X_{dm} = resize\ (X_{cm}, [a, b])$

$X_{dfc} = \sum_{j=1}^{N} X_{cm} * w_{n+1} + X_{mp} * H(k)$

from (10& 13)

$X_{da} = X_{fc} * F_m(i)$

End loop

Model2 = $\{X_{dm}, X_{dm}, X_{dFc}, X_{da}\}$

Procedure: IAD

Create Pipeline:

For i = 1: length(models):

$Model_x = w_n * model1 + loss_{fn}(model2)$

End loop

End pipeline
End procedure

4.7 *Evaluation Parametric*

The proposed design with the parametric features is considered and utilized to realize the importance classification values based on the six terms indicating the best performance of the design before and after optimization. The values are listed below:

1. Sensitivity
2. F1-score
3. Recall
4. Precision
5. Accuracy
6. Specificity.

To represent the such parametric a confusion matrix is implemented to realize the design has high performance based on the above six parametric. Generally, the matrix for disease classification are 2x2 matrices consisting of four elements as listed below:

1. True Positive
2. True Negative
3. False Positive
4. False Negative

The TP, TN are sample values observed after predictions which represents that a person has heart disease or not. While FP, FN are the values for misdiagnosed for patients who have heart disease or not. The formulations for the parametric are represented below:

$$Accuracy = \frac{TP + TN}{FP + TP + TN + FN}$$

Precision for the positive case is determined by:

$$Precision = \frac{TP}{TP + FP}$$

Other important features such as sensitivity and specificity are given by:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100$$

$$Specificity = \frac{TN}{TN + FP} * 100$$

6. RESULTS AND DISCUSSION

In this section, the proposed design and existing results are verified based on the list of algorithms with respect to the literature survey provided. We indicate the different parametric values on each algorithms improvising any changes that have been effectively, realized to implicate how the prediction of heart disease is possible. We also consider how each of the algorithms are effective before and after optimization based on the parametric values as mentioned above.

The results from the first phase of design from the figure (6)-(7) and (8) represents the overall Heat-map graphs visualized with balanced and un balanced cases for 22 features considered. The implementation work on the design indicates the overall sample size of 0.49 million samples of patients were considered based on the Balanced approach. To balance the data, we have utilized SMOTE functionality for the design indicate the overall sample size from 0.249 million to 0.49million. As per the visualization the we could observe that the data values for the correlation matrix for 22x22 are converted using corr() function. Figure 6 represents the unbalanced condition from 90-10 percent to 50-50 percent indicating the correct balanced form of data cleaning operations as mentioned in implementations.

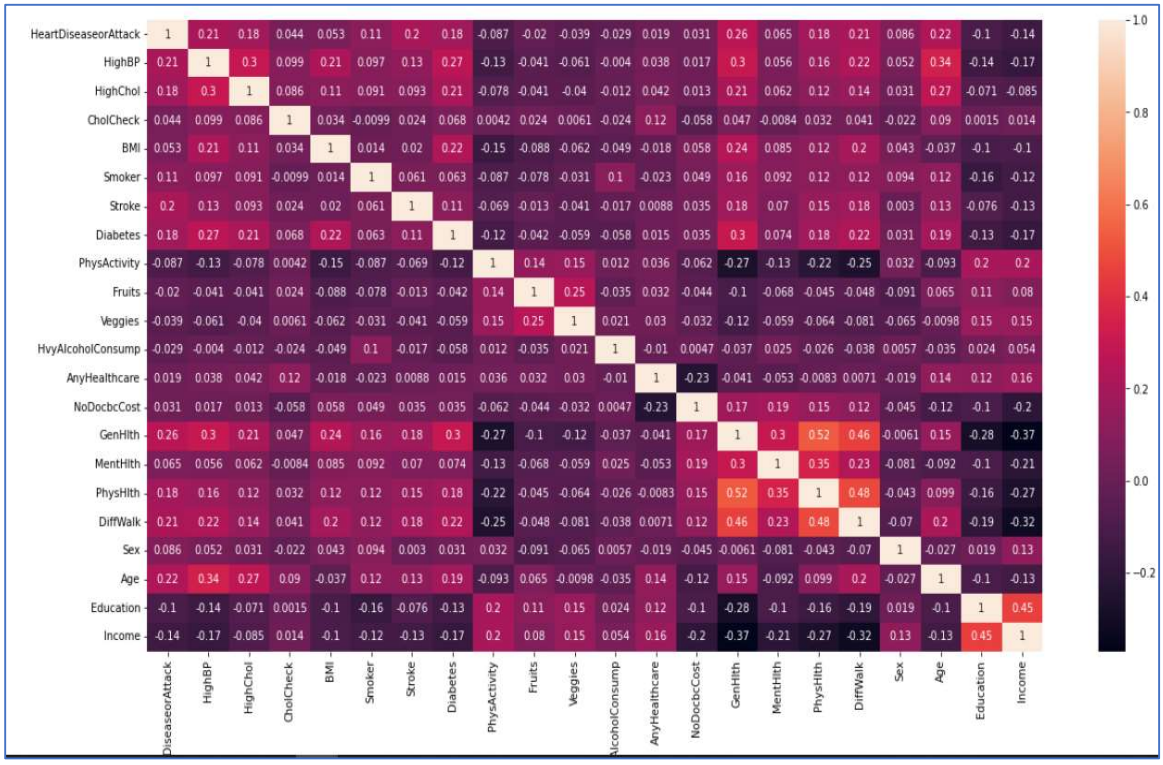

Figure 6: Representing the full Heat Map for 22 Columns for Classification.

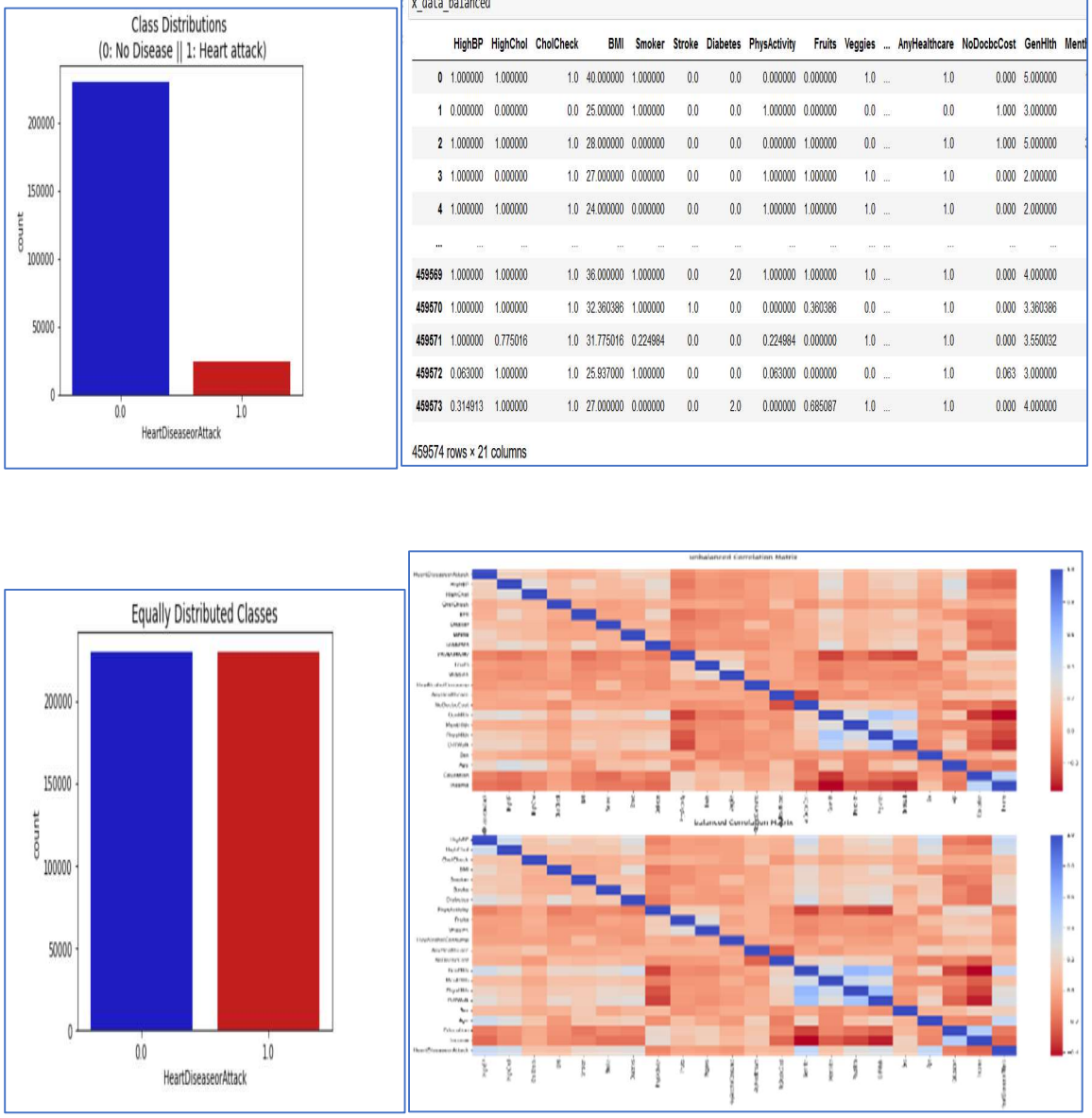

Figure 7: Representing The Unbalanced Condition From 90-10 Percent To 50-50 Percent

While in figure 8, we utilize the Seaborn plots representing the histograms of each feature with respect to classification labels. The X-axis values represent the overall labels in the design while the Y axis represents the total sample size for each feature indicated with ('0' and '1') labels.

Journal of Theoretical and Applied Information Technology
15st September 2026. Vol.104. No.17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

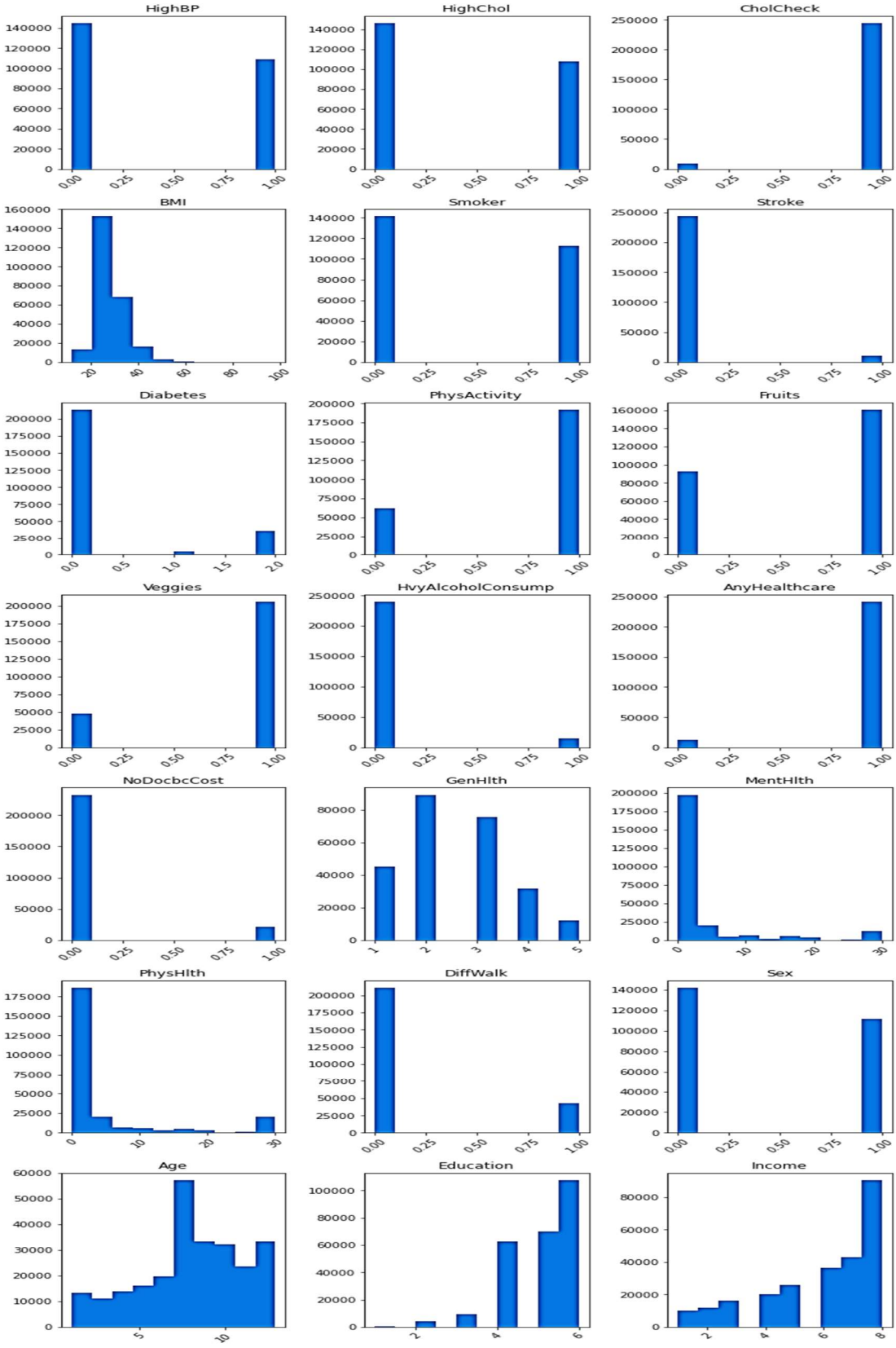

Figure 8: Represents The Seaborn- Pair Plots for Each Sampled Responses With Balanced Cases.

While in second phase of the design indicates the overall plotting of ROC plot and loss Vs accuracy for training and testing models indicating the two algorithms before and after optimization. The optimization technique with SVD+PCA analysis have proven better and low loss when compared to existing SOA architectures. For the below figure 8 representing the accuracy and loss graphs for training and testing accuracies for algorithm - IADM and IAD. The overall estimated loss observed with an average value of 0.24 for validation while the overall accuracy is 95.9%. While for IAD model we have observed similar characteristics with an average value of 90.8%.

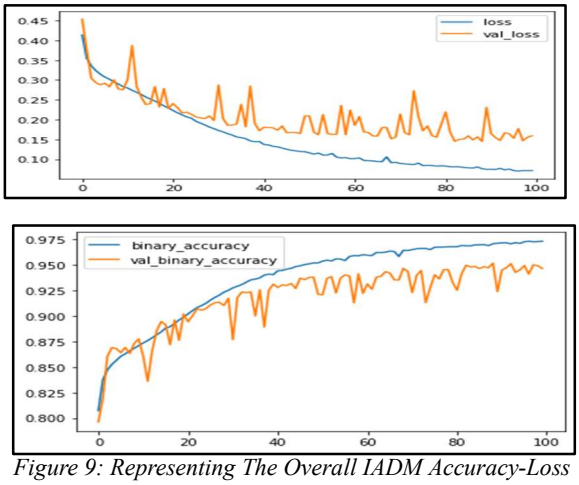

Figure 9: Representing The Overall IADM Accuracy-Loss Graph.

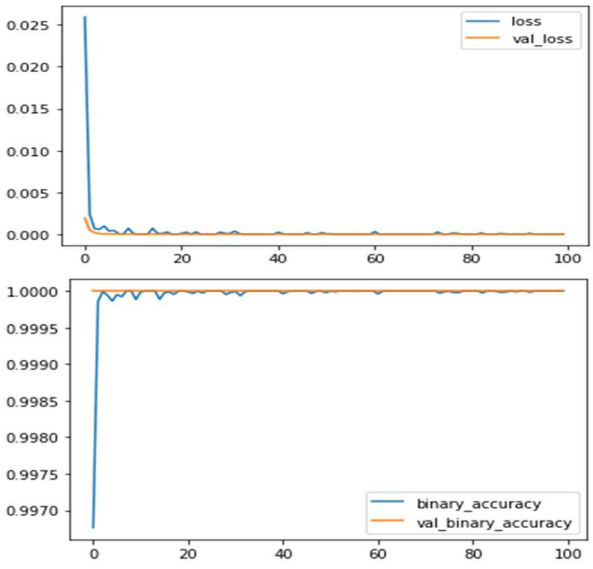

Figure 10: Representing The Overall IAD Accuracy-Loss Graph

We have implemented the proposed IADM and IAD models with 100 epochs criteria as shown above with validation accuracy and loss. These values of consistency are observed for the optimized cases indicating best accuracy with Deep learning model design. The hybrid design on IADM and IAD have utilized with ADAM optimizer with the batch size of 32 samples on the data length of 0.49 million. we could observe from the figure 9 the overall optimized accuracy of the proposed IADM and IAD is 100%. While for existing SOA architectures are similar when compared with the same optimization values. The prediction 200-2000 samples for the disease prediction have been considered indicating the correct cases and true cases of the samples based on the confusion plots. All such values are estimated based on the performance metrics as mention before and after optimization in table 2-4 below.

The performance comparison across various machine learning and deep learning models in table-2 for heart disease prediction reveals significant differences in classification effectiveness. Traditional models like Logistic Regression (LR) show lower sensitivity (52.86%) and specificity (47.43%), indicating limited capability in distinguishing between heart disease and non-disease cases. Random Forest Classifier (RFC) and Decision Trees (DT) significantly outperform LR with sensitivities around 95%, but exhibit lower F1-scores (~85.6–86.3%) compared to deep learning approaches. KNN achieves high recall (99.2%) but suffers from low specificity (36.7%), leading to imbalance in classification.

Ensemble-Hybrid models improve performance across all metrics, achieving an F1-score of 93.6%. Notably, the **proposed IADM model** achieves **98.09% sensitivity** and an F1-score of **94.8%**, closely matching or surpassing state-of-the-art (SOA) LSTM-based models. The enhanced **IAD model** further improves sensitivity to **98.42%**. While LSTM and LSTM+Hybrid models perform well, with F1-scores above 94% and precision nearing 92.5%, IADM maintains a strong balance between **recall (98.1%)** and **precision (91.7%)**, showcasing its robustness on noisy or incomplete datasets. Overall, IADM and IAD models

demonstrate competitive or superior performance against SOA methods, confirming the effectiveness of attention-based interpolation in improving diagnostic reliability.

Table 2: Representing The Overall Performance Metric Before Optimization with Existing and Proposed Approaches.

Algorithms	Sensitivity	Specificity	F1-Score	Recall	Precision
LR [10]	52.86	47.43	78.8	81	76.7
RFC [7]	95.3	94.65	85.6	87.2	84.6
DT [23]	95.01	95.23	86.3	87.4	85.3
KNN [2-5]	63.3	36.7	87.5	99.2	78.2
ENSEMBLE-HYBRID [6]	94.24	92.91	93.6	94.2	93
Proposed Hybrid-CNN	**98.09**	**91.11**	**94.8**	**98.1**	**91.7**
Proposed IDDM-CNN	**98.42**	**90.35**	**93.5**	**97.5**	**91.2**
LSTM (SOA) [30]	98.35	96.23	94.8	98.4	92.5
LSTM+HYBRID (SOA) [31]	98.65	94.173	95.4	95.17	90.73

Table 3: Representing The Overall Performance Metric After Optimization with Existing And Proposed Approaches.

Algorithms	Sensitivity	Specificity	F1-Score	Recall	Precision
LR [10]	100	100	100	100	100
RFC [7]	100	100	100	100	100
DT [23]	100	100	100	100	100
KNN [2-5]	100	100	100	100	100
ENSEMBLE-HYBRID [6]	99.84	99.91	99.26	99.9	99.73
Proposed IADM	**100**	**100**	**100**	**100**	**100**
Proposed IAD	**100**	**100**	**100**	**100**	**100**
LSTM (SOA) [30]	99.5	99.3	99.88	99.84	99.65
LSTM+HYBRID (SOA) [31]	99.95	99.17	99.96	99.83	99.997

Table 4: Representing The Overall Comparative Table for Existing and Proposed Approaches.

Algorithms	Accuracy (Existing) Without Optimization	Proposed (Accuracy) With Optimization Class
Lr [10]	**78.7**	**100**
Rfc [7]	**93.7**	**100**
Dt [23]	**85.6**	**100**
Svm [2]	**82.4**	**100**
Knn [2-5]	**86.4**	**100**
Ensemble-Hybrid [6]	**94.5**	**100**
Proposed IADM	**94.6**	**99.86**
Proposed IAD	**90.96**	**92.45**
Lstm (Soa) [30]	**91.85**	**98.76**
Lstm+Hybrid (Soa) [31]	**96.54**	**99.15**

Table 3 highlights the performance metrics of various algorithms after optimization, where most models—including LR, RFC, DT, KNN, and the proposed IADM and IAD—achieve perfect scores (100%) across sensitivity, specificity, F1-score, recall, and precision. Even the state-of-the-art LSTM and LSTM+Hybrid models perform exceptionally well, with metrics exceeding 99%, indicating that optimization significantly enhances all models' predictive reliability, especially in handling classification balance and precision-recall trade-offs.

Table 4 provides a before-and-after comparison of model accuracy, showing substantial improvements post-optimization. Traditional machine learning models such as LR, DT, SVM, and KNN exhibit a sharp jump from moderate accuracy (78–86%) to 100%. Similarly, ensemble models and LSTM-based architectures also see notable gains. The proposed IADM model improves from 94.6% to 99.86%, while IAD increases from 90.96% to 92.45%, demonstrating the effectiveness of the attention-based interpolation strategy. Overall, both tables confirm that optimization significantly enhances model performance, with the proposed methods delivering results comparable to or better than existing state-of-the-art approaches.

7. CONCLUSION

The main contribution of this study is the development of an attention-guided interpolation-based learning framework that jointly addresses incomplete, uncertain, and heterogeneous clinical information in heart disease prediction. By integrating interpolation-based estimation with attention-guided feature learning, the proposed framework is designed to estimate incomplete clinical information and identify informative relationships among patient attributes, thereby reducing the potential adverse effects of missing and heterogeneous clinical data on prediction. The study further provides a systematic comparative evaluation against conventional machine learning, hybrid learning, and state-of-the-art deep learning approaches using the selected publicly available clinical datasets. The experimental results demonstrate the effectiveness of the proposed framework and provide evidence of its predictive robustness and consistency across datasets with different characteristics.

Overall, this work contributes to the development of robust data-driven approaches for heart disease prediction by shifting the focus from prediction under predominantly complete clinical information toward learning under more realistic incomplete and heterogeneous data conditions. The findings provide a computational basis for future data-driven clinical decision-support applications. However, the proposed framework is not intended to provide direct clinical diagnosis or replace physician decision-making, and clinical deployment would require further validation using larger, multi-institutional, longitudinal, and real-world clinical datasets. Future research will therefore investigate external validation, model explainability, uncertainty quantification, robustness under varying missing-data mechanisms, and prospective clinical evaluation.

REFERENCES

[1]Danish Hamid, Syed Sajid Ullah, Jawaid Iqbal, Saddam Hussain, Ch. Anwar ul Hassan, Fazlullah Umar, "A Machine Learning in Binary and Multiclassification Results on Imbalanced Heart Disease Data Stream", *Journal of Sensors*, vol. 2022, Article ID 8400622, 13 pages, 2022.

[2]S. Chitra and V. Jayalakshmi, "Analyze the Medical Threshold for Chronical Kidney Diseases and Cardio Vascular Diseases using Internet of Things," 2021 4th International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, 2021, pp. 189-193.

[3]T. Kavitha, S. Hemalatha, S. Mownika, G. M. Priya and J. K. Vigashini, "Review on Cardio Vascular Disease Prediction Using Machine Learning," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-4.

[4]C. Singla, B. K. Sahoo, R. Kaur, P. Sahu, H. Singh and U. Kaur, "Prediction and Classification of Cardio Vascular Diseases using Ensemble Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1665-1667.

[5]G. U. Santosh Kumar, T. V. Rajini Kanth, S. V. Raju and S. Malyala, "Advanced Analysis of Cardiac Image Processing Using Hybrid Approach," 2021 International Conference on Advances in Electrical, Computing,

Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2021, pp. 1-6.

[6]B. Anishfathima, R. Vikram, S. R. T, M. Sri Vishnu and C. Venumadhav, "A Comparative Analysis on Classification Models to predict Cardio-vascular disease using Machine Learning Algorithms," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, pp. 259-264.

[7]Saravana Selvan, S. John Justin Thangaraj, J. Samson Isaac, T. Benil, K. Muthulakshmi, Hesham S. Almoallim, Sulaiman Ali Alharbi, R. R. Kumar, Sojan Palukaran Thimothy, "An Image Processing Approach for Detection of Prenatal Heart Disease", *BioMed Research International*, vol. 2022, Article ID 2003184, 14 pages, 2022.

[8]G. Kumar Sahoo, K. Kanike, S. K. Das and P. Singh, "Machine Learning-Based Heart Disease Prediction: A Study for Home Personalized Care," 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), Xi'an, China, 2022, pp. 01-06.

[9]G. Shobana and S. N. Bushra, "Prediction of Cardiovascular Disease using Multiple Machine Learning Platforms," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2021, pp. 1-7.

[10]G. M. Shree Raksha, R. Hegde, M. N. Shivani, P. S. Shrinidhi, M. M. Thashwin Monnappa and S. M. Soumyasri, "A Novel Technique for Prediction of Cardiovascular Disease," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2022, pp. 1-5.

[11]K. S. Archana, B. Sivakumar, Ramya Kuppusamy, Yuvaraja Teekaraman, Arun Radhakrishnan, "Automated Cardioailment Identification and Prevention by Hybrid Machine Learning Models", *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 9797844, 8 pages, 2022.

[12]N. Sarmiladevi, K. Basker, R. Tamilarasu and L. Nivetha, "An Efficient Detection System For Cardio Vascular Disease Using Machine Learning Algorithms," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 951-955.

[13]P. Anuradha and V. K. David, "Feature Selection and Prediction of Heart diseases using Gradient Boosting Algorithms," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 711-717.

[14]Abdullah Alqahtani, Shtwai Alsubai, Mohemmed Sha, Lucia Vilcekova, Talha Javed, "Cardiovascular Disease Detection using Ensemble Learning", Computational Intelligence and Neuroscience, vol. 2022, Article ID 5267498, 9 pages, 2022.

[15]A. K. Shukla, "Prediction of Heart Disease through Machine Learning Algorithms and Techniques," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-6.

[16]J. Fontecave-Jallon, A. Haouas and S. Tanguy, "Abdominal cardiovascular sound recording and analysis using cardio-microphones," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 820-823.

[17]R. Wang et al., "Calcium Activation of Parvalbumin Neurons Induced by Electrical Motor Cortex Stimulation," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 2353-2356.

[18]P. M., S. V. M., J. A. and A. Mayan, "Cardiovascular Disorder Prediction using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1665-1670.

[19]C. Saravanabhavan, T. Saravanan, D. B. Mariappan, N. S, D. Vinotha and K. M. Baalamurugan, "Data Mining Model for Chronic Kidney Risks Prediction Based on Using NB-CbH," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 1023-1026.

[20]S. Patil and S. Bhosale, "Assessing Feature Selection Techniques for Machine Learning Models using Cardiac Dataset," 2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA, 2022, pp. 126-130.

[21]D. Lakshmi Padmaja, B. Sai Sruthi, G. Surya Deepak and G. K. Sri Harsha, "Analysis to Predict Coronary Thrombosis Using Machine Learning Techniques," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 21-27.

[22]Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 8387680, 11 pages, 2021.

[23]K. Nithyakalyani, S. Ramkumar, S. Rajalakshmi and K. A. Saravanan, "Diagnosis of cardiovascular disorder by CT images using Machine learning technique," 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), Chennai, India, 2022, pp. 1-4.

[24]Abdul Saboor, Muhammad Usman, Sikandar Ali, Ali Samad, Muhmmad Faisal Abrar, Najeeb Ullah, "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, vol. 2022, Article ID 1410169, 9 pages, 2022.

[25]L. Deng, Y. Zhang, Z. Li and K. Wu, "An Adaptive Estimation of Ultrasound Transit Time-based Local PWV in Carotid Artery using Particle Swarm Optimization Algorithm," 2021 IEEE International Ultrasonics Symposium (IUS), Xi'an, China, 2021, pp. 1-4.

[26]T. Venkat Narayana Rao, G. Vishal Sai, P. H. Vardhan Reddy, S. Harsha Bandarupally and C. Nikhil, "SHANE – Smart HeartRate Analysis and Notification System for Emergencies," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 1104-1107.

[27]R. Kumar M, G. K. E, D. R and D. K. M, "Detection Of Arrhythmia Using Machine Learning(Heart Disease) And ECG," 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, 2022, pp. 1-4.

[28]M. Revathi, A. B. Godbin, S. N. Bushra and S. Anslam Sibi, "Application of ANN, SVM and KNN in the Prediction of Diabetes Mellitus," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), Chennai, India, 2022, pp. 179-184.

[29]B. S. K. Jayasudha, P. N. Sudha, K. Keshav and N. Ramesh, "Ensemble learning as a prerogative method of predicting mortality of patients with cardiovascular diseases," 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2021, pp. 01-05.

[30]K. Divya Vani, “Heart Disease Prediction using Long Short-Term Memory (LSTM)Deep Learning Methodology”, International Journal of Research Publication and Reviews Vol (2) Issue (8) (2021), Page 1076-1081.

[31]Rath, A., Mishra, D., Panda, G. (2021). LSTM-Based Cardiovascular Disease Detection Using ECG Signal. In: Mallick, P.K., Bhoi, A.K., Marques, G., Hugo C. de Albuquerque, V. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 1317. Springer, Singapore.