DESIGN AND DEVELOPMENT OF FEATURE ENGINEERING MODEL FOR CVD MULTI-DIMENSIONAL DATASETS STANDARDIZATION

Sangeetha V[1], Syed Muzamil Basha[2*]

[1,2]School of Computer Science and Engineering, REVA University, Bengaluru, India

Email: sangeethacs025@gmail.com[1], muzamilbasha.s@reva.edu.in[2*]

ABSTRACT

Cardio Vascular Diseases (CVD) are most common medical abnormalities in modern times. According to World Health Organization (WHO), the CVD related deaths spot is higher and increased due to chronical reasons such as unhealthy lifestyle, food habits, geographical changes and comorbidities from genetic history. The studies are focused on one particular causes as it trends to eliminate the comorbidities influencing the secondary causes. Hence a large unbalanced prediction and independent datasets are created and customized. In this research, the authors have aimed to propose dataset standardization on CVD. The proposed standardization is based on attribute interdependency mapping and indexing. The attribute interdependency from one influencing parameter is aligned and coordinated with secondary attributes from relationship/dependency evaluation. Further, the dependency mapping is simplified by layering and customizing the data attributes. The multi-dimensional CVD datasets extracted in this process is mapped and tracked from feature engineering process. The dataset standardization of CVD is novel and first of its kind in data analytics and processing.

Keywords: Cardio Vascular Diseases (CVD); dataset standardization; machine learning; CVD; data processing; multidimensional datasets.

1.0 INTRODUCTION

Cardiovascular disease (CVD) encompasses a spectrum of pathologies affecting the cardiovascular system and ranks as a prominent global health concern. The historical evolution of CVD elucidates the progression of medical knowledge. Ancient civilizations demonstrated recognition of cardiac disorders, while the 19th century initiated the establishment of systematic cardiology. Substantial strides in CVD diagnosis and treatment emerged during the 20th century, exemplified by the creation of the electrocardiogram and landmark investigations like the Framingham Heart Study, elucidating key risk factors. Surgical advancements, including coronary artery bypass grafting and angioplasty, marked the late 20th century. In the 21st century, cutting-edge imaging modalities and the implementation of personalized medical approaches have enriched our comprehension and therapeutic strategies for CVD. Nevertheless, persistent challenges arise, notably the escalating prevalence of CVD attributed to sedentary lifestyles and dietary choices. As such, the battle against CVD retains its pivotal status within modern healthcare.

Cardiovascular Diseases (CVD) represent the prevailing health disorders in contemporary society, a fact substantiated by the World Health Organization (WHO) and corroborated by data [1] indicating a concerning rise in CVD-related mortality, which has surged from 2.26 million in 1990 to an alarming 4.77 million in 2020. This surge in mortality underscores the heightened gravity of the issue. Notably, the urban population with a metropolitan lifestyle is experiencing a significant increase in CVD mortality, reaching a striking 13.4%. The primary aim of this research is the establishment of a standardized dataset for the systematic evaluation and validation of chronic CVD occurrences. The research manuscript is constructed around the integration of three distinct and independent CVD datasets, thereby creating a comprehensive, multidimensional CVD dataset. This dataset's ambit extends to encompass multiple layers of attribute dependencies and features, ultimately facilitating early-stage CVD normalization and risk assessment.

2.0 RELATED WORK

Cardiovascular Diseases (CVD) and the corresponding challenges they present have become prominent focal points of contemporary research. One of the primary limitations hindering such research efforts is the scarcity of comprehensive datasets. Typically, available datasets are tailored to address specific facets of CVD, neglecting the intricate interconnections among attributes and their collective impact on CVD decision-making processes.

The fundamental objective of this survey is to delve into the multitude of parameters and attributes that play a role in CVD validation, with the ultimate aim of establishing a robust foundation for informed decision support. In a noteworthy contribution, a study detailed in [1] has been introduced, focusing on CVD classification and prediction through the utilization of Machine Learning (ML) and Deep Learning (DL) techniques. This research relies on extensive data mining efforts to unearth meaningful patterns that aid in CVD classification. Yet, a significant research challenge persists, involving the accurate identification of CVD occurrences within the dataset. Addressing this concern, another study outlined in [2] has devised an enhanced preprocessing methodology, offering improved predictive capabilities and CVD classification. This approach represents a critical step toward more accurate and reliable CVD decision support systems.

Machine learning models play a pivotal role in bolstering the reliability of decision support systems concerning Cardiovascular Disease (CVD) by enabling the meticulous mapping of CVD parameters and attributes while facilitating the extraction of interconnected features. This study, referenced as [3], contributes significantly to the discourse on advancing CVD modeling and monitoring through the integration of Machine Learning (ML) techniques. Notably, it addresses the persistent challenge of model overfitting and randomization, a hurdle mitigated by the work presented in [4][6]. In this endeavor, an enhanced validation technique is introduced, involving the construction of a multi-layer decision tree that takes into account the interdependencies among attributes. This approach facilitates the comprehensive indexing of attributes in conjunction with their correlated features, ultimately enhancing the accuracy of CVD prediction and classification. It is crucial to note that the validation process detailed in [4] is confined to a refined dataset, leading to the subsequent exploration presented in [5][6][7]. This latter study introduces an innovative approach aimed at customizing CVD prediction through a rigorous risk estimation framework. By doing so, it extends the frontiers of CVD research towards more tailored and effective decision support systems[8].

Datasets serve as a linchpin in the intricate realm of cardiovascular disease (CVD) research, and their role is inherently interwoven with the quality of results and predictive capabilities generated by the models. However, a prevalent limitation in the current landscape of CVD studies, often exemplified by references [9] and [10], is the utilization of datasets that are predominantly unidimensional or narrowly focused on specific attributes, such as isolated data on blood pressure, cholesterol levels, or a handful of isolated factors. This one-dimensional approach to dataset construction neglects the complex and interconnected nature of the various parameters associated with CVD, resulting in models that may lack comprehensive predictive power. Breaking away from this conventional approach, a pioneering study documented in [11] delves into the potential of cross-domain-based CVD prediction. This paradigm shift involves the integration of datasets from other medical domains, notably datasets pertaining to kidney-related health. By doing so, it enables the validation of the intricate interdependencies among attributes and influential variables that contribute to CVD prediction. This cross-domain approach heralds a more holistic understanding of CVD, as it acknowledges the multifaceted nature of the disease and its interactions with various aspects of health.

Further, the research outlined in [12] and [13] [14] builds upon the foundation laid by this cross-domain strategy. These studies highlight the importance of cross-domain attribute mining and pattern extraction as essential components of CVD decision-making processes. Such approaches permit a more nuanced analysis of CVD by uncovering intricate patterns and relationships between variables from different domains, ultimately enhancing the precision of CVD prediction and decision support systems. [15] [16] Thus, the pressing need for multi-dimensional and customized standardized CVD datasets becomes abundantly clear in the context of higher-order research and analysis. Such datasets act as a cornerstone for pioneering advancements in CVD research, affording researchers the tools necessary to comprehensively explore the multifaceted nature of CVD, extract meaningful patterns, and deliver more precise predictions and insights.

3.0 PROBLEM STATEMENT

The Cardio Vascular Diseases (CVD) datasets are retrieved, archived and managed via independent data repositories. This scenario makes the overall processing and computation a challenging task. Consider the scenario of computing CVD occurrence of age group (25 – 45 years), the typical considerations are bound to Electronic Health Records (EHR) based attributes (finite set) as $\{A_{DM1}, A_{DM2}, A_{DM3} \dots A_{DMn}\}$ and making a scenario of $(D_M \epsilon A_{DMi})_0^n$ and thus attributes associated to external source datasets (i.e.) (D_S) is not compatible as $(\forall A_{DMi} \sqsubseteq D_M) \& (D_M \notin D_S) \Longrightarrow (D_S \notin A_{DMi})$. This theory of analysis complicates the processing as shown in below figure. 1.

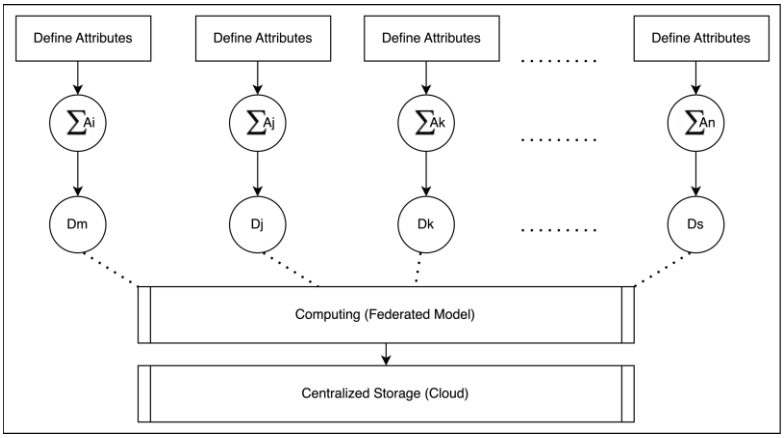

Fig. 1: Independent computation of datasets with isolative attribute extraction and storage in centralized server (Cloud)

Thus from Fig. 1, we can draw a conclusion for reflective processing and customization of datasets from centralized servers. Since cloud is a centralized platform for storage, the processing of feature engineering is aligning multi-source datasets is challenging task. Thus, the resultant computation of these independent datasets can generate multi-dimensional CVD datasets and standardization.

4.0 MATERIALS AND MATHEMATICAL MODEL

4.1 Methodology

The proposed system is aimed to process and develop a multi-dimensional dataset on CVD. Thus the initial representation includes a series of datasets $(S_1, S_{2,}, S_3 \ldots.)$ and aggregated in a single source of operation (i.e.) data aggregator. Typically, the dataset collected from these sources are independent and have not direct coordination's from one attribute to another. The data aggregator collects datasets and provides a primary repository dataset for computation. This repository is treated as pre-trained source of this proposed system under initial cycle of training and labeling. The process is accompanied with attribute extraction and dependency mapping of the pre-trained CVD datasets such as $(A_1, A_{2,}, A_3 \ldots.)$ are extracted. Each attribute value is summarized to the maximum isolation and hence the dependencies are mapped as shown in Fig. 2.

The dependency mapping and attribute indexing is taken up as a next phase of computation. The dependency $(D_1, D_{2,}, D_3 \ldots.)$ is extracted from 1st layer to nth layer of mapping phases termed as layer mapping. Hence the resultant mapping values are customized and pre-trained labels are computed towards data standardization process. The generation of each layer and individual participations is demonstrated in Fig. 3 (Classification diagram). The phase includes a detailed multi-source dataset collection, dataset aggregation, attribute extraction unit and feature engineering phase for the generation of nth layer multi-dimensional CVD datasets.

4.2 Data Aggregator

The proposed system is developed with the objective to extract and evaluates the dataset from multiple sources and repositories. The CVD datasets are typically aligned and are codependent to include the relevance of parameters and attributes. The CVD datasets used in this research study are MIMIC – III (i.e.) Medical Information Mart for Intensive Care datasets, Framingham Heart Study (FHS), Cleveland Heart Diseases (CHD) datasets and Arthrodesis risk in communities (ARIC) datasets for the process of interdependency mapping and multi-dimensionality dataset creation. The inclusiveness of these datasets are further corresponsive to the interdependency attribute and feature mapping for developing the complex interdependency maps of CVD datasets in general. Fig. 4 represents the attribute feature corelationship mapping. The attribute dependencies and complexity is evaluated with the minimal indexing ratio of each attribute with its parallel association to create a feature parameter. Thus according to Fig. 4, the computation of each independent dataset has a unique and unshared parameters of attributes. The attribute ratio of these are indexed to form an aggregator sum of interdependent datasets. The dependency formulation on the grounds of (L_1, L_2) aggregator is limited to occurrence and causes complexity in evaluation. Typically, the dataset values are processed on a single stream of

computational indexing as represented in Fig. 4 and further extending towards dependency processing stage for multi-dimensional CVD dataset creation.

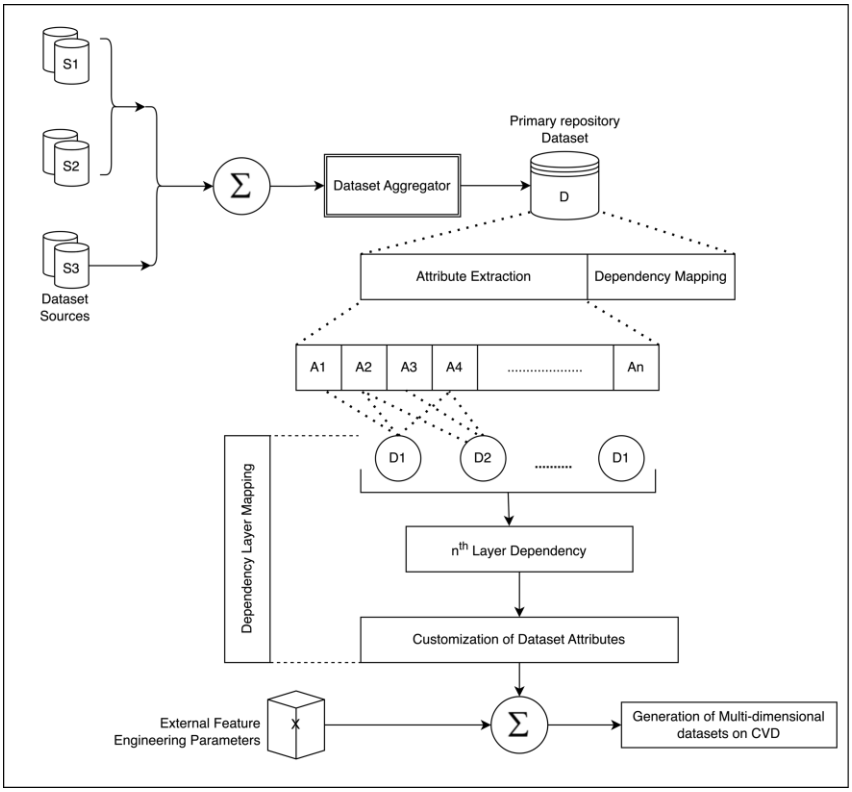

Fig. 2: Proposed system block diagram

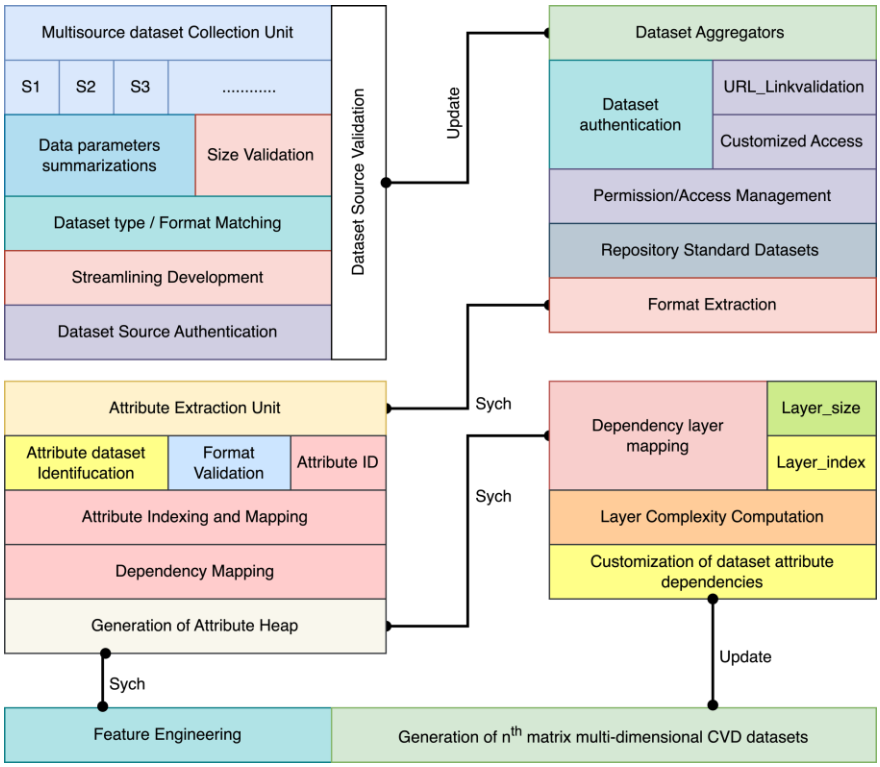

Fig. 3: Classification and stack-flow representation of multi-dimensional CVD dataset generation.

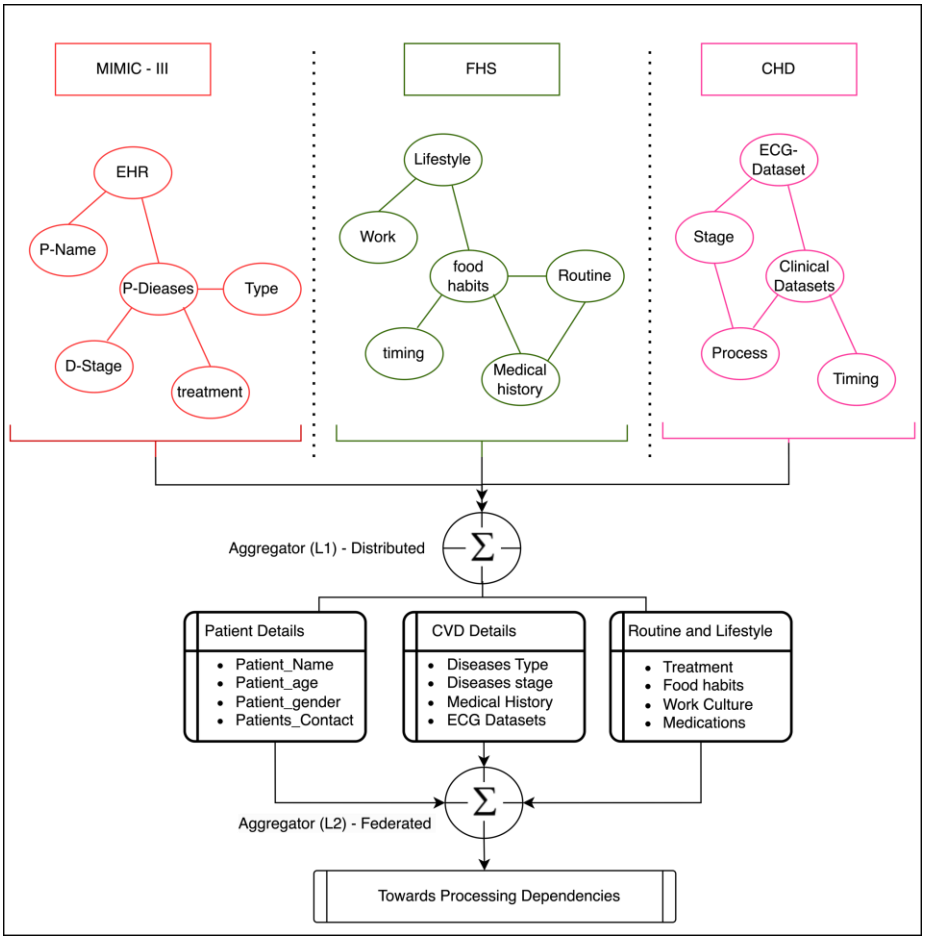

Fig. 4: Relationship representation diagram and index mapping of attributes towards CVD multi-dimensional dataset creation.

4.3 Dependency Processing

The dependencies are typically associated with the formulation matrix evaluation of two or more independent datasets in computation. Consider the datasets as (D_X) with $(D_1, D_{2,}, D_3 \ldots.)$ are contributing datasets in the (D_X) computation. Typically, $(\forall D_X \Longrightarrow D_i/D_i \in \Sigma A_i)$ where (ΣA_i) are primary dependencies matrix associated with (D_X). On considering the scenario of primary (L_1) dependency extraction, the $(D_i \Longrightarrow (A_1, A_2, A_3 \ldots A_n) \sqcup (A_{11}, A_{12}, A_{13} \ldots A_{1n}) \sqcup (A_{21,}, A_{22}, A_{23} \ldots) \ldots)$ at the ratio mapping. Typically the $\Sigma(A_{i1}, A_{i2}, A_{i3} \ldots)$ are the associated attributes from each independent dataset as $(\Sigma A_{ij} \in D_i)$ such that $[\forall A_{ij} \Longrightarrow \ni (D_i \cap D_j)]$ for mutual ratio extraction.

On further understanding, the dependency breaking can be represented as shown in Eq. 1, where (ΣD) is the dependency matrix of given datasets with multiple sources. Hence the computational values of (ΣD) is dependent on $(\Delta\, A_{(i,j)})$ associated to (D_i) at the given time (t), this can be customized as shown in Eq. 2.

$$\Sigma D = \lim_{n \to \infty} \left(\frac{\partial (D_i)}{\partial t} * \iint_0^n \frac{\partial \left(\Delta A_{(i,j)} \right)}{\partial t} \right) \tag{1}$$

$$\Sigma D = \prod_n^\infty \sum_{i=1}^n \left(\left[\frac{\partial (D_i)}{\partial t} \right] \Longrightarrow \left[\iint_0^n \frac{\partial \left(\Delta A_{(i,j)} \right)}{\partial t} \right] \right) \tag{2}$$

$$\therefore \Sigma D = \frac{1}{t} \left\{ \prod_n^\infty \sum_{i=1}^n \left([\partial (D_i)] \Longrightarrow \left[\iint_0^n \partial \left(\Delta\, A_{(i,j)} \right) \right]_{(i,j)}^n \right) \right\} \tag{3}$$

Thus, according to Eq. 3, the time $(t)^{-1}$ is extracted as a constant parameter on processing the dataset (D_i) and attribute dependencies $\left(A_{(i,j)}\right)$ for a given ratio of volume. Typically, the coordination of $(D_i) \Longrightarrow (\Delta t)_0^n$ and hence add indexes to the processing parameter. According to Eq. 3, the dependencies indexes (I) are shown in Eq. 4.

$$I = \left[A_{(i,j)}\right]_{(i,j)\Rightarrow n} * log_n(D_i) \tag{4}$$

with passing epoch, the indexing volume (I) is added to the individual attribute ratio $\left(A_{(i,j)}\right)$ with (D_i) existence. If the association is reflected in dual parameters dataset, for instance $\left[\left(D_i \longrightarrow D_j \longrightarrow D_k\right) \Longrightarrow \Delta\, A_{(i,j)}\right]$ at generalized processing of parameter attributes. Though $\left(\forall\, \Delta\, A_{(i,j)} \in D_i, D_j, D_k\right)$ the indexing parameters is strengthen on (D_i) as origin value of dataset also termed as "master node" of $\left(\Delta\, A_{(i,j)}\right)$ at given primary instance (t) as shown in Eq. 5 and Eq. 6.

$$I_1 = Occ\left[A_{(i,j)} * \sigma t\right] \oplus log_n(D_i) \tag{5}$$

$$I_2 = Occ\left[A_{(i,j)} * \sigma t\right] \oplus log_n(D_i) \tag{6}$$

the collective representation is as shown in Eq. 7.

$$I_{Occ} = \begin{matrix} D_i : if\ \left[A_{(i,j)}\right] \Longrightarrow Count = 0 : Call(case\ 1) \\ D_j : if\ \left[A_{(i,j)}\right] \Longrightarrow Count = 1 : Call(case\ 1 + loop) \end{matrix} \tag{7}$$

thus the summarization can be drafted as Eq. 8 from Eq. 7.

$$\Sigma(I_{Occ}) = argmin\left\{\left(\Delta A_{(i,j)}\right) \oplus \frac{\partial(D_i) \cup \partial(D_j) \cup \partial(D_k) \ldots}{\partial x}\right\} \tag{8}$$

$$\Sigma(I_{Occ}) = argmin\left\{\left(\Delta A_{(i,j)}\right) \oplus \sum_{K=1}^{n} \frac{\partial(D_i)_k}{\partial x}\right\} \tag{9}$$

Thus, the reflective index of two binary information from (I_{Occ}) in Eq. 9 summarizes the existence of attribute and the dependencies layer (L_i) with respect to dataset (D_i). typically, the values of these datasets are internally computed and a multi-dimensional CVD dataset is extracted.

4.4 Customization of CVD multidimensional datasets

The process of CVD dataset customization includes the refinement of dependency ratio as demonstrated in Eq. 9. The dataset attributes customs include the factor of stacking most influencing parameters of datasets (S), this includes as shown in Eq. 10.

$$\Sigma(I_{Occ}) = \Delta argmin\ \{\Sigma(I_{Occ})_0^n\} \tag{10}$$

The customization results in generating a multi-dimensional CVD datasets as shown in Fig. 4 respectively. The propagation stream of data coordination and alignment assures the corelationship between the extracted attributes is interdependent on formulated indexing matrix. Typically, the mapped attributes are further associated and customized with reference to number of participating datasets (D_X) and associated layers of attribute count. The outcome of this research is to generate a reliable and customized CVD datasets for regression and testing of novel CVD related scenarios. The justification is followed with retaining the basic input and dataset origin points for collaborative learning via transfer learning models on public domain. The variation and fragmentation of datasets origin results in collaborative outcomes of attributes associations and dataset monitoring.

5.0 RESULTS AND DISCUSSIONS

The proposed system is developed on a generic computational environment using NVidia GPU servers blocks for computation and operations. Typically, the side-lined computation and operations as endues (i.e.) data origin servers are regular compute with i5 silver line processor connected via master-slave representation as shown in block diagram (Fig. 1). The experimentation is aligned on universal CVD datasets (i.e) MIMIC – III, FHS, CHD and ARIC for higher accuracy and real-time compatibility. The proposed framework has generated a higher order MIMIC – III and FHS combined dataset aggregation for relatively higher order of attribute dependency extraction. The outcome of this research is to streamline the dataset (CVD) with respect to monitoring repositories for resolving newer and customizable features associated in CVD such as unnatural CVD attacks pattern extraction, studying habits and behavior based on geographical populations and its influence in CVD and much more. The realistic possibilities of this research is novel multi-dimensional dataset extraction and mapping with respect to the real-time CVD datasets (single dimensional) to generate or explore larger possibilities.

Table.1: Computation of dataset v/s attribute dependencies (Indexes)

Dataset_name	Dependency attribute (%)	Non-Dependency attribute (%)	Interdependency attribute (%)
MIMIC – III	46.32	50.36	3.32
FHS	68.17	19.41	12.42
CHD	52.11	41.17	6.72
ARIC	29.64	68.72	1.64

According to Table.1, the relevance of dataset based inference of dependency, interdependency and non-dependencies is computed. The validation is concluded based on the observation from existing CVD based single dimension model design. According to outcome patterns as shown in Table. 2 of three prominent CVD datasets (i.e) MIMIC – III, FHS and CHD, we have extracted the dimensional representation of dependency matrix and interdependency matrix with respect to the indexing ratio of given datasets on layer of indexing. Typically, the indexing is resultant of multiple values associated with the dataset-attributes on particular function mapping. The corelationship of these attributes are to demonstrate they associations with nearest alike attribute for decision making. For instance, the value of TP (True Positive) and FN (False Negative) depicts the influence attribute mapping and clarity of attributes association.

Table.2: Computational model for layer based attribute dependency indexing

Dataset	Layers	Dependency matrix (%)	Interdependency matrix (%)	Indexing ratio	
				TP (%)	FN (%)
MIMIC – III	L1	46.32	3.32	41.11	58.89
	L2	49.67	11.17	56.42	43.58
	L3	49.68	18.11	58.52	41.48
	L4	52.34	19.66	58.42	41.48
FHS	L1	68.17	12.42	66.48	33.52
	L2	69.41	14.11	68.14	31.12
	L3	70.91	14.82	68.88	31.12
	L4	71.20	15.67	70.66	29.34
CHD	L1	52.11	6.72	48.11	51.89
	L2	53.66	8.42	49.72	50.28
	L3	53.67	8.49	49.93	50.06
	L4	53.67	8.49	49.93	50.06

6.0 CONCLUSION

The proposed technique has made significant strides in the realm of Cardiovascular Disease (CVD) research, as evidenced by successful applications and findings on multi-dimensional CVD datasets. The primary outcome of this research effort is the creation of a pioneering and unique class of multi-dimensional CVD datasets, a novel development in the field. These datasets are meticulously customized to align with the specific inferences and requirements of both end-users and researchers in the CVD domain. Importantly, this dataset is open to the

broader scientific community and can be readily harnessed for a multitude of applications, particularly within the domains of behavioral studies and pattern evaluation. The datasets have exhibited a remarkable improvement in the accuracy of mapping and indexing the intricate correlation matrix that exists among various attributes and their respective features, a critical advancement that enhances the precision and comprehensiveness of CVD analysis. Looking ahead, there is immense potential for these datasets to be amalgamated and universally applied in open domain validation across a range of datasets. This interoperability opens doors for collaborative research efforts and the further refinement of CVD insights on a broader scale.

REFERENCES

1. Huffman, Mark D., et al. "Incidence of cardiovascular risk factors in an Indian urban cohort: results from the New Delhi Birth Cohort." *Journal of the American College of Cardiology*, Vol. 57, No. 17, 2011, pp. 1765-1774.
2. Swathy, M., and K. Saruladha. "A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques." *ICT Express*, Vol. 8, No. 1 , 2022, pp. 109-116.
3. Louridi, Nabaouia, Meryem Amar, and Bouabid El Ouahidi. "Identification of cardiovascular diseases using machine learning." *2019 7th mediterranean congress of telecommunications (CMT)*. IEEE, 2019.
4. Moradi, Hamed, et al. "Recent developments in modeling, imaging, and monitoring of cardiovascular diseases using machine learning." *Biophysical Reviews*, Vol. 15, No. 1, 2023, pp. 19-33.
5. Balakrishnan, M., et al. "Prediction of Cardiovascular Disease using Machine Learning." *Journal of Physics: Conference Series*. Vol. 1767. No. 1. IOP Publishing, 2021.
6. Patil, Prasadgouda B., P. Mallikarjun Shastry, and P. S. Ashokumar. "Machine learning based algorithm for risk prediction of cardio vascular disease (Cvd)." *Journal of critical reviews*, Vol. 7, No. 9, 2020, pp. 836-844.
7. Marbaniang, Ibashisha A., Nurul Amin Choudhury, and Soumen Moulik. "Cardiovascular disease (CVD) prediction using machine learning algorithms." *2020 IEEE 17th India Council International Conference (INDICON)*. IEEE, 2020.
8. Al'Aref, Subhi J., et al. "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging." *European heart journal*, Vol. 40, No. 24, 2019, pp. 1975-1986.
9. Saikumar, K., and V. Rajesh. "A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset." *International Journal of System Assurance Engineering and Management*, 2022, pp. 1-17.
10. Guarneros-Nolasco, Luis Rolando, et al. "Identifying the main risk factors for cardiovascular diseases prediction using machine learning algorithms." *Mathematics*, Vol. 9, No.20, 2022, pp. 2537.
11. Matsushita, Kunihiro, et al. "Incorporating kidney disease measures into cardiovascular risk prediction: development and validation in 9 million adults from 72 datasets." *EClinicalMedicine*, Vol. 27, 2020.
12. Peng, Mengxiao, et al. "Prediction of cardiovascular disease risk based on major contributing features." *Scientific Reports*, Vol. 13, No. 1, 2023, pp. 4778.
13. Wei, Xi, et al. "Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model." *Expert Systems with Applications*, Vol. 219, 2023, pp. 119648.
14. LK, Sowmya Sundari, et al. "COVID-19 outbreak based coronary heart diseases (CHD) prediction using SVM and risk factor validation." *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE, 2021.
15. Kumar, S. Sreedhar, et al. "Unstructured Oncological Image Cluster Identification Using Improved Unsupervised Clustering Techniques." *Computers, Materials & Continua*, Vol. 72, No. 1, 2022.
16. Ahmed, Syed Thouheed, et al. "IMPROVING MEDICAL IMAGE PIXEL QUALITY USING MICQ UNSUPERVISED MACHINE LEARNING TECHNIQUE." *Malaysian Journal of Computer Science*, 2022, pp. 53-64.