IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 29, 2021, accepted December 26, 2021, date of publication December 28, 2021,
date of current version January 5, 2022.

*Digital Object Identifier 10.1109/ACCESS.2021.3139123*

# Performance Enhancement of Predictive Analytics for Health Informatics Using Dimensionality Reduction Techniques and Fusion Frameworks

**SHAEELA AYESHA, MUHAMMAD KASHIF HANIF, AND RAMZAN TALIB**
Department of Computer Science, Government College University, Faisalabad 38040, Pakistan
Corresponding author: Muhammad Kashif Hanif (mkashifhanif@gcuf.edu.pk)

**ABSTRACT** Predictive analytics has become an essential area of research in health informatics. The availability of multi-source and multi-modal data in healthcare has made the disease prediction, diagnosis, and medication process more effective and reliable. However, the analysis and decision making have become challenging task, particularly when data is in multiple formats and from different sources. In this study, different frameworks have been proposed to handle multi-nature data at different levels for predictive analytics. Dimensionality reduction techniques have been applied to extract relevant features to enhance the analysis. To improve the performance of predictive analytics at different fusion levels, the potential benefits of multi-modal data have been discussed. Moreover, notable improvement in prediction accuracy has been observed through experimental evaluation of the proposed frameworks. Furthermore, the issues which have been found during dimension reduction and fusion approaches have also been highlighted.

**INDEX TERMS** Dimensionality reduction, feature extraction, data fusion, feature fusion, machine learning, high dimensional data.

## I. INTRODUCTION

Over the past few decades, the evolution of Information and Communication Technology in healthcare has brought revolutionary changes in data collection tools and techniques. Healthcare is a data-intensive field generating a massive volume of data every day. Health informatics systems are making a significant contribution to transform healthcare from a data-intensive to a data-rich field. An intense increase in the data size and dimensions used for different applications demands novel approaches to make effective use of data to enhance the performance of these systems [1], [2].

Health informatics systems play an essential role in hospitals and healthcare centers. These systems need to be modernized as they face many issues in acquiring relevant and complete data [3], [4]. Various issues have been found in health informatics data such as uncertainty, imperfection, heterogeneity, inconsistency, redundancy, high dimensionality and representation create problems for the fusion process [5]. The performance of disease management and prediction systems is often adversely affected due to data quality issues [6]. To overcome these issues, it becomes essential to remove discrepancies from the data to discover valuable hidden information for analysis and decision making.

Data belonging to different sources and modalities have different features that provide different information about disease and patient condition. Thus feature fusion of multi-source and multi-modal data can improve the accuracy and reliability of results as compare to single modality [7]. Moreover, relying on the data from a single source or modality can be challenging to discriminate different diseases and classification of severity level. The fusion of multi-source and multi-modal data has become common in recent years. Multi-modal systems not only consider different aspects of disease management, but also allows for missing data imputation, quality-aware fusion, and improve the precieved experience.

A variety of techniques can be used to analyze current and historical information to make accurate predictions. The predictive analytics approaches allow the maintenance of actions based on changes in the parameters and factors that can affect disease diagnosis. However, the quality of input data has a significant impact on the performance of the prediction system [8], [9]. In HDD, multiple parameters belonging to one entity may be collected to diagnose disease.

The associate editor coordinating the review of this manuscript and approving it for publication was Sedat Akleylek.

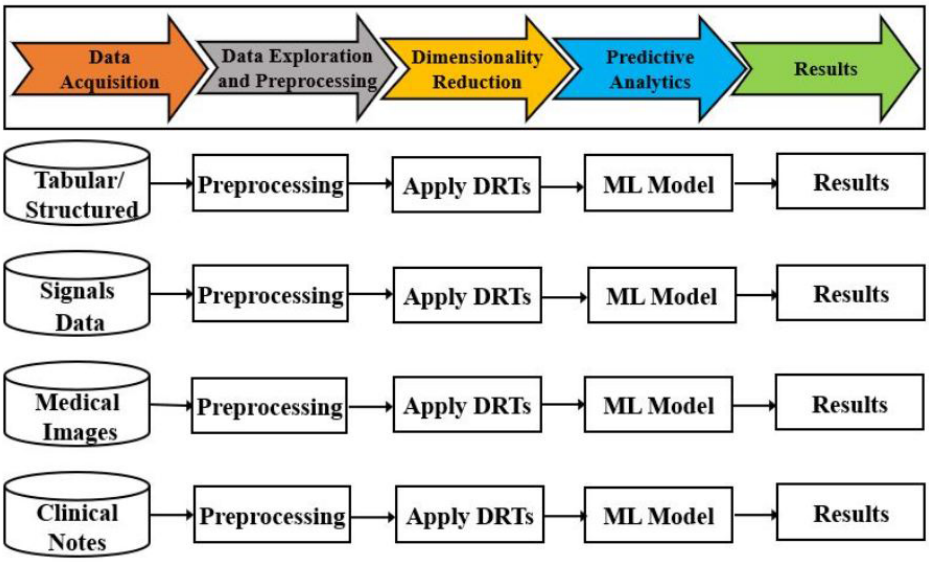


**FIGURE 1.** Architecture of classical uni-modal approach for predictive analytics. Here, predictive analytics is performed for data $x_i$ retrieved from each source $S_i$ separately. First, data from each source $S_i$ is preprocessed to clean the data $x_i$. Then DRT transforms the data $x_i$ into lower dimension $y_i$. Finally, predictive analytics model is applied according to type of data to attain the results.

HDD often holds redundant, sparse, missing, noisy, and irrelevant features [10]–[12]. Moreover, analysis of HDD and diverse nature data is challenging for decision-making and disease prediction.

The predictive analytics approaches allow the maintenance of actions based on such changes in the parameters and factors that affect disease diagnosis. A variety of techniques can be used to analyze current and historical information to make accurate predictions. However, the quality of input data has a significant impact on the performance of the prediction system [8], [9]. In some situations, hundreds or even millions of measurements belonging to one entity may be collected to diagnose disease. In such a situation, generally, the issue of HDD is raised. HDD often holds redundant, sparse, missing, noisy, and irrelevant features [10]–[12]. Moreover, analysis of HDD and diverse nature data is challenging for decision-making and disease prediction; only relevant features are required.

Many Dimensionality Reduction Techniques (DRTs) have been successfully applied in the literature to extract only precise and relevant features for analysis and disease prediction. Dimensionality Reduction (DR) can be performed via feature selection and feature extraction [13]. Feature selection creates a subset of features using correlation analysis or weighting methods [14]. Feature extraction transforms the original HDD into a low-dimensional representation by eliminating the redundant features [15]. Feature extraction can be performed using linear approaches [16], such as Principal Component Analysis (PCA), Independent Component Analysis (ICA). And nonlinear approaches [17], such as Kernel PCA (KPCA), Isomap, and Self Organizing Map (SOM), etc. Different DRTs can be applied for different types of data (e.g., image, text, signals) [11]. Moreover, different researchers have applied different machine and deep learning models, such as Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF) [18], [19], Clustering [5], [20], and Neural Networks (NN) [21] for disease classification and prediction [3], [22].

At present, most of the frameworks proposed in the literature are using a single modality for predictive analytics (e.g., the prediction of disease) (Figure 1). For example, Electrocardiogram (ECG) signals have been used for arrhythmia detection [23]. The problem with this approach is that ECG signals and ECG reports may not be available simultaneously. But, other information such as symptoms of heart disease, risk factors information, and demographic data may be available, which can be helpful for disease prediction. An ECG-based prediction system that uses ECG signals only can limit the scope and need modification in such a situation. Similar issues have been found with other disease prediction and management systems developed to work with one modality at a time [2], [24]. Some studies proposed fusion frameworks such as for readmission prediction [25], disease prediction which are specific for the fusion of multi-modal or multi-source data at a time [18], [26]. These systems offer better results than unimodal based system. There is need for a mechanism than can utilize multi-source, multi-modal, and multi-nature data simultaneously to develop dynamic and efficient disease prediction and management systems.

Several techniques have been proposed to combine data from multiple sources and modalities in recent years, which may hold similar and different data formats used for predictive analytics [18], [21], [26]–[28]. These techniques include data integration, data fusion, feature fusion, knowledge fusion, and multi-sensor fusion [4], [21], [26]. There is need to develop a system that can fuse multi-source,

multi-nature, and multi-modal data to improve the performance of disease diagnostic and prediction systems [5]. Multi-source fusion increases the reliability and availability of data. However, different sources may hold data in different modalities (e.g., formats). When the objective is to fuse (combine) diverse nature multi-source data, it becomes more suitable to apply multi-source fusion. But multiple sources may contain redundant features, so there is a need to fuse features belonging to different modalities. Feature fusion models can combine feature sets belonging to two or more modalities to acquire distinct, relevant, and precise features. Many studies explored different levels of fusions [3], [29], [30]. Selection of required fusion level needs a clearly defined objective for combing various features belonging to different sources in the form of different modalities and vice versa. The resulting (combination of) features belonging to multiple sources and modalities is expected to generate improved prediction performance compared to individual modality or data belonging to a single source [5].

This study proposed different fusion frameworks to combine the data from multiple sources and modalities for predictive analytics to overcome the issues mentioned above. Moreover, knowledge and decision fusion frameworks have been introduced to improve decision-making. It becomes challenging to use a particular DRT or a common classifier or Machine Learning (ML) model for multi-nature and multi-modal data [11], [19], [31]. To resolve this challenge, suitable DRTs can be applied to extract the most relevant, reliable, and precise features to improve the performance of the classifier or the ML model. Ultimately, to enhance the performance of predictive analytics for health informatics.

This study makes the following contributions:

- Multi-modal, multi-source data, feature, knowledge, and decision fusion frameworks have been proposed to combine the data from multiple sources and modalities to improve predictive analytics.
- For considering the contribution of different DRTs and ML models, we explored different DRTs and ML models for the proposed fusion frameworks.
- Key issues belonging to DRTs and fusion frameworks for predictive analytics have also been highlighted.

The present study is organized as follows: Section II presents the related work. Fusion frameworks are proposed in section III. Section IV discusses the results of the proposed frameworks. Section V explores the key challenges that limit the application of DRTs and fusion approaches. Finally, section VI concludes the study.

## II. REVIEW OF LITERATURE

Several multi-source, multi-modal, and knowledge fusion frameworks have been introduced in the literature to improve the performance of disease prediction systems. This section provides a brief review of different fusion frameworks suggested in the literature for handling high-dimensional, multi-source, and multi-nature health informatics data. According to reviewed literature three common fusion levels have been found. Data fusion approaches such as integration and fusion have been used to fuse multi-source or multi-modal data. Feature fusion approaches have been defined to combine features from multiple sources and modalities. Feature selection and extraction methods can be applied for feature fusion (e.g., to combine features belonging to different sources and modalities). Decision fusion approaches aims to combine the results of different classifiers using major voting, min-max score, etc.

### A. MULTI-MODAL DATA FUSION

In literature, most of the studies proposed fusion frameworks perform multi-modal fusion of different medical signals (e.g., ECG, EEG, etc.) [7] or imaging modalities (e.g., MRI, PET, CT, or CXR) [32], [33]. To resolve the issues of multi-source data fusion a genetic algorithm was proposed to select data from source having optimal information [34]. In a recent study, Muhammad *et al.* [7] conducted a brief review of various studies from 2014 to 2020 that perform multi-modal fusion of medical signals to develop efficient health management systems. A multi-modal data fusion strategy was proposed to detect the progress of Alzheimer's [35]. Author, applied Spatial Group ICA was used to reduce the dimensionality of imaging modalities and Canonical Correlation Analysis (CCA) was used to fuse the features of fMRI and sMRI. However, disease classification task was performed with SVM and recurrent convolutional network.

In study [36], authors proposed a multi-modal fusion framework to combine the features of image and speech for smart health monitoring. A multi-sensor data fusion methodology was developed for blood pressure assessment system for CVD management in ambulatory care [37]. For different health informatics applications sensor data fusion frameworks have been proposed in different studies [1], [38], [39].

For the collection of multi-source data from different medical sensors to develop a personalized health management system, Korzun and Meigal introduced a fusion framework based on semantic links to combine data from various sensors to improve the performance [1]. Nweke *et al.* reviewed different data fusion and features fusion strategies [31]. A taxonomy of multi-sensor and multi-view features and data fusion has been developed for human activity recognition. Both classical DRTs and automatic deep learning methods were used for multi-sensor and feature fusion based on multi-view fusion approaches.

In another study, for the prediction of human personality, Kampman *et al.* [8] proposed a multi-modal fusion framework using major voting (ensemble) method to combine text, audio, and video data that can be used in healthcare. Majumder and Pratihar proposed a multi-sensor fusion via fuzzy clustering for the prediction of heart diseases [21]. Vijayasarveswari *et al.* introduced a multi-phase feature selection approach for cancer prediction [26]. Zhang *et al.* introduced a multi-modal fusion framework based on Local Linear Projection (LLP) [40]. Khan *et al.* proposed a feature

fusion strategy to improve the prediction and classification accuracy of Coronary Artery Disease (CAD) [41].

### B. MULTI-MODAL FEATURE FUSION

During multi-modal feature fusion data belonging to different modalities, sources, and sensors either homogenous or heterogeneous are collected and combined. For instance, to enhance the heart disease management system via multi-modal fusion, [42] applied PCA to reduce the dimensionality and genetic algorithm for classification. For the prediction of heart disease at an early stage, a multi-sensory-based data and feature fusion framework have been proposed by Muzammal *et al.* [18]. They extracted time domain and frequency domain features after preprocessing of collected data. Multi-sensor and multi-modal features were fused to develop a decision support system for heart patients. They used a fog cloud-based. A review of fusion implementation in health informatics has been found in [43]. In this study, authors discussed different fusion level used to fuse imaging and Electronic Health Records (EHRs) data using deep learning techniques. For automatic diagnosis of myocardial infarction via ECG data, Wang *et al.* [19] introduced a multi-feature fusion approach. The proposed approach performed feature fusion based on DRTs, statistical features, and entropy features. Extracted features were classified using RF. The proposed approach achieved a higher disease classification accuracy than state-of-the-art approaches.

In clinical practice, the collection of multi-modal medical imaging data has become common. To extract relevant and reduced features and to overcome the issue of data sparseness, different DRTs such as Sparse PCA, PCA, ICA, and CCA have been applied by Yang *et al.* at preprocessing phase [44]. After preprocessing, they fused extracted features for brain disease diagnosis. [45] proposed a multi-modal feature fusion framework for the detection and classification of Arrhythmia and Myocardial Infarction (MI) from ECG signals data collected from MIT-BHI dataset using CNN layers for feature extraction and SVM as a classifier. For the detection of Atrial Fibrillation (AF), Shi *et al.* proposed a feature fusion framework. Discriminative CCA was applied to reduce the dimensionality of CMRI data [46].

To enhance the prediction accuracy of heart disease prediction systems for smart health monitoring, Ali *et al.* proposed a feature fusion framework to combine the structured and unstructured data using the deep learning approaches [47]. In another study, Ali *et al.* [4] proposed an analytical engine for efficient prediction of disease based on multi-source data. They apply different data mining strategies to reduce the context-aware dimensionality of data and train and test different ML models for disease prediction.

Terms used in the table: Multi-source Fusion (MSF), Multi-modal Fusion (MMF), Feature Fusion (FF) Decision Fusion (DF), Reduced Feature set Fusion (RFF), Knowledge Fusion (KF), Information Fusion (IF) Tabular = Tab, Image = Img, Text = Txt, and Signal = Sig.

### C. DECISION FUSION

Decision fusion is considered as the highest fusion level. Decision fusion was implemented in various application of health informatics. For the development of efficient disease detection system, the authors in [55] preprocessed two different input modalities separately and than fused them at classifier level (fuse the results of different classifiers). When fusing multi-modal and HDD, completeness and quality of data have a significant impact on fusion [5]. Deng *et al.* introduced a classifier (decision) fusion framework based on feature selection to overcome the issue of data incompleteness and fusion of high dimensional and multi-modal medical imaging modalities [14] used for the classification of Alzheimer's disease. The proposed feature selection-based classifier fusion approach achieved higher accuracy with incomplete datasets.

Smirnov and Levashova presented a review of knowledge fusion patterns reported in different studies [30]. According to this study, three kinds of fusion was performed, e.g., knowledge fusion for the knowledge stored in repositories (e.g., accumulation of knowledge, problem-solving, multi-source fusion, searching, concept, attribute or domain fusion), knowledge fusion among knowledge workers (e.g., knowledge sharing, problem-solving, decision making, or distribution of knowledge), knowledge fusion among knowledge workers and knowledge repositories (e.g., analysis and problem solving). In a recent study, Tariq *et al.* proposed a COVID-19 disease prediction and resource management system where decision level fusion of multiple classifiers was performed [56]. To enhance the classification accuracy of the COVID-19 patients a decision fusion method was introduced [54].

### D. MULTI-LEVEL FUSION

Some studies performed fusion at different stages /levels to develop a dynamic and efficient systems for health informatics. For instance, For the efficient diagnosis of heart disease, Hassan *et al.* [49] introduced a multi-stage fusion (e.g., feature and decision fusion) using generative model and multi-variate process control method. The proposed multi-stage fusion architecture used shared and separated ICA to overcome the issue of high dimensionality, incompleteness, heterogeneity, due to multi-modal and non-normalized features in data. A brief review study for different data fusion (level) for healthcare enhancement was provided in [57].

For the identification of different diseases (Alzheimer's and Cancer), based on multi-modal clinical data, Viswanath *et al.* [48] proposed a fusion framework for knowledge representation of multi-modal data via decision, kernel, and low-dimensional representations. Knowledge fusion based on weighting criteria was performed using direct fusion, co-association matrix fusion, and structural fusion. They presented results by combining multi-modal interpreters. DRTs were used for data and feature fusion at different levels.

**TABLE 1. Summary of fusion frameworks used predictive analytic in health informatics.**

| Reference | Purpose | Fusion Level | | | | | | | Modalities | | | | DRTs | Classifiers | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MSF | MMF | FF | RFF | KF | DF | IF | Tab | Img | Txt | Sig | | | |
| [48] | Alzheimer's and Cancer disease Prediction | - | - | - | - | + | - | - | - | + | - | - | PCA | SVM | - |
| [14] | Alzheimer's disease | - | + | - | - | - | - | + | - | + | + | - | - | Polynomial kernel | 96.08 |
| [49] | Diagnosis of heart disease | - | - | + | - | - | + | - | + | - | - | + | Separated and shared ICA | SVM | 93.83 |
| [8] | Personality prediction for health management | - | - | - | - | - | - | + | - | + | + | - | CNN | CNN | 0.9033 |
| [21] | Heart disease prediction | + | - | - | - | - | - | - | + | - | - | - | Clustering with fuzzy approach | SVM, NN, and Version Space SVMs(VSSVMs) | 100 |
| [50] | Survey of fusion strategies for healthcare | - | + | - | - | + | - | - | - | + | + | - | DNN | ML models | - |
| [31] | Human activity detection for healthcare management | + | - | - | | - | - | + | - | + | - | + | PCA | SVM, DT,KNN, and LR | - |
| [44] | Brain disease classification | + | - | - | - | - | - | - | - | + | - | - | PCA, sPCA, CCA+ sCCA, and Parallel ICA | No | p =90 |
| [19] | Detection of Myocardial infraction | - | - | + | - | - | - | - | - | - | - | - | PCA | RF | 99.71 |
| [18] | Heart disease prediction | + | - | - | - | + | - | - | - | - | - | - | - | RF and Kernel RF | 98 |
| [41] | CAD detection and classification | - | - | - | + | - | - | - | + | - | - | ++ | No | KNN, SVM | 88.4 |
| [45] | Arrhythmia and MI classification | + | - | - | + | - | - | - | - | - | - | + | CNN | SVM | 99.2 |
| [46] | Heart disease detection | - | - | + | - | - | - | - | - | + | - | - | DCCA | - | - |
| [51] | COVID-19 diagnosis | - | - | - | - | + | - | - | - | - | - | - | PCA | SVM, CNN | 94.7 |
| [52] | COVID-19 detection | - | - | - | - | - | + | - | - | + | - | - | | KNN, SVM, ANN | 96.9 |
| [53] | COVID-19 prediction | - | - | + | - | - | + | - | - | + | - | - | CNN | CNN | 89.84 |
| [54] | COVID-19 classification | - | + | + | - | + | - | - | - | + | - | - | CNN | SVM, KNN, RF | 99.2 |

To explore the issues of the ageing population and the spectrum of pandemics in the modern age, Cai *et al.* [50] presented a survey of the top 10 countries publishing research articles focusing on data-driven health management systems. They explored various approaches and applications used for multi-modal data and knowledge fusion in the data-intensive healthcare domain. A fusion framework for two modalities (image and text data) has been presented to perform data and knowledge fusion for a Clinical Decision Support System (CDSS). SMOTE method was used for feature fusion. In another study, Zhang *et al.* discussed multi-view data fusion and feature fusion strategies using feature selection approaches [58]. For monitoring blood pressure using ECG sensors, Smirnov and Levashova proposed a multi-level fusion model that can perform multi-sensor fusion and information fusion for the development of a predictive model [30].

To improve the COVID-19 diagnostic accuracy [53] proposed feature fusion and decision fusion methodology. Multiple CNN architectures were used to extract features and feature level fusion was performed to combine the features. Different classifiers were trained and tested on the fused features set. The Majority Voting method was used for decision fusion to attain the optimal results for medical recommendations to control the pandemic. Attallah *et al.* proposed a diagnostic system for distinguishing COVID-19 and non-COVID-19 cases [51]. The system was trained and tested using CT images, where the CT image features were extracted with four pretrained deep CNN models, and than were fused for training SVM classifier. The authors experimented with different fusion strategies to investigate the impact of feature fusion on the diagnostic performance.

To overcome the challenges of uni-modal data for emotion recognition, Jiang *et al.* proposed a multi-modal information fusion approach [59]. The proposed approach was used to extract relevant features from multi-modal data (e.g., ECG signals, visual, audio, text data) and fuse them for emotion recognition and health monitoring of patients. For dimensionality reduction and feature extraction, classical DRTs, such as Linear Discriminant Analysis (LDA) and some deep learning approaches, such as CNN, RNN, and DBN, were applied. They used feature and decision fusion. DRTs were applied for feature fusion. For decision fusion, results of different classifiers were combined via weighted sum to obtain the collective decision. Other studies introduced multi-level fusion such as fusion techniques suggested for health monitoring via wearable sensors [60], fusion of text and images for COVID-19 detection [52], decision and feature fusion for COVID-19 data sets [53], feature and data fusion [61] etc.

Table 1 summarizes the reviewed literature along with key findings concerning various fusion frameworks and the significance of DRTs to improve the performance of various ML models. It has been explored that most of the studies focused on one or two types of fusion (e.g., multi-modal feature and or knowledge fusion) for a particular situation, which makes their scope limited. According to the reviewed literature, to the best of our knowledge, no study has been proposed which support fusion frameworks for multi-source, multi-modal, multi-nature data and feature fusion to combine reduced feature sets and decision or knowledge fusion frameworks to combine the results of different ML models for predictive analytics. Moreover, another contribution of this study, we applied various combinations of DRTs and ML models to select the best approach used for different fusion frameworks.

## III. PROPOSED FUSION FRAMEWORKS FOR PREDICTIVE ANALYTICS

This section illustrates the structure and functionality of the proposed fusion frameworks in detail. The steps of the uni-modal data processing pipeline are used as a baseline for the development of disease prediction models and systems (Figure 1). According to the existing studies, a uni-modal data processing pipeline can work with one modality or one type of data at a time and cannot handle the heterogeneous and high-dimensional health informatics data for predictive analytics (e.g., disease prediction). In this study, different fusion frameworks have been proposed to process heterogeneous (e.g., multi-nature, multi-source, and multi-modal) and high dimensional health informatics data to develop an efficient and reliable disease prediction system (Figure 2 to 6).

These steps include data acquisition from multiple sources belonging to multiple modalities having different data types (e.g., medical images, text, signals, and tabular data). Various operations are performed at the data preprocessing and exploration phase to prepare data for further processing. During the dimensionality reduction phase, different DRTs are applied to reduce the dimensionality of data and to extract precise features for analysis. Then reduced feature set is forwarded to the predictive analytics phase, where different ML models are applied to predict disease efficiently. Finally, medical treatments can be recommended based on the prediction and classification of disease.

Moreover, due to variation in the significance of multiple features for solving the specific issue, not all the features of the data collected from multiple sources are useful for analysis and decision making. To extract relevant features from different modalities, DRTs according to the type and nature of data can be applied. Extracted features belonging to different sources and modalities can be combined (fused) for analysis. These reduced feature sets hold more precise and relevant features for disease prediction, hence, improving the performance of the disease prediction system.

For example, data collected from multiple sources (e.g., patient conversation with physician, medical diagnostic devices, lab test reports, other online and offline sources, medical research, surveys, etc.). The data $\boldsymbol{X}$ belonging to different sources $\boldsymbol{S}$ is combined in a uniform format for analysis using fusion frameworks. Each dataset $x_i$ can have $f_k$ features, where $1 \leq k \leq d$. The fusion of multi-nature datasets may result in duplication of data and features. The data $\boldsymbol{X}$ is represented in the form of a matrix $\boldsymbol{Z}$. The rows and columns of matrix $\boldsymbol{Z}$ are obtained from $x_i$ and $f_k$, respectively.

In this study, a collection of DRTs are applied to eliminate redundant features. The new collective source obtained after fusion can offer a uniform representation of data. Data from multiple modalities $\boldsymbol{M}$ is denoted by $m_j$ where $j = 1, 2, \cdots, q$. Modalities indicate different formats of data found in healthcare, such as structured (e.g., patient demographic information, hospital visit history, and billing information), semi-structured (e.g., clinical notes), and unstructured (e.g., signal, images). In this study, we categorized these modalities (data types) as tabular/structured, signals, images, and text formats. These modalities can hold information, such as patient demographic information, risk factors, and medical reports (e.g., lab tests and ECG reports). Each modality $m_j$ stores data in the form of features that are represented as matrix $\boldsymbol{Z}$. The preprocessing phase $\boldsymbol{P}$ for each modality $m_j$ is $p_j$ to develop a standardized representation of features. These features may comprise of patient information gathered from multiple sources $\boldsymbol{S}$ and modalities $\boldsymbol{M}$. Next, DRT are applied to transforms the HDD $\boldsymbol{Z} \in \mathbb{R}^{l \times d}$ having $d$ dimensions and $l$ rows into lower dimensions $\boldsymbol{Y} \in \mathbb{R}^{l \times r}$ where $r < l < d$ in an ideal case.

Features extracted for analysis via DRTs will be represented as $f_t$ where $1 \leq t \leq r$. Moreover, the reduced feature set for each modality $m_j$ is represented by $h_u$ where $1 \leq u \leq r$. The reduced representations after applying DRTs can be quite different from original data. For example, after applying PCA on a dataset, reduced data will be presented in the form of Principal Components (PCs) based on the function used for calculation of these components, e.g., covariance or variance, etc. The reduced feature set $h_u$ belonging to multiple modalities $m_j$ can be fused and sent to the selected ML models/ classifiers $c_w$. Results of different ML models can be combined to make a collective decision which is also called decision fusion. Different frameworks have been proposed in the following subsections, which represent the fusion of data and features at different phases of a standard data processing pipeline used for predictive analytics.

### A. MULTI-SOURCE DATA FUSION FRAMEWORK

Data belonging to a single source may not hold sufficient information for effective medical decision-making [62]. To overcome this issue, data from multiple sources can be acquired and combined using a multi-source data fusion framework (Figure 2). The main objective of the multi-source data fusion framework is to increase the accuracy and reliability of data to improve the accuracy of the results. According to Figure 2, the proposed multi-source data fusion framework combines data from multiple sources and preserves them in EHRs. During the data acquisition phase, different datasets $x_i$ belonging to multiple sources $S_i$ are specified for analysis.

$$\boldsymbol{X} = \sum_{i=1}^{n} x_i \tag{1}$$

where $x_i \in S_i$

At the preprocessing phase $\boldsymbol{P}$, various preprocessing operations such as imputation missing values, outlier detection and removal, and class balance are performed to remove anomalies. The preprocessed data $X_p$ is obtained by imputing missing values and balancing the class. The preprocessed data may be greater then or equal to original data.

$$\boldsymbol{X}_p \geq X \tag{2}$$

The dimensionality reduction phase reduces the dimensionality of data and extracts relevant features for analysis. DRTs are applied according to the type and nature of data. DRT is applied to obtain reduced data $\boldsymbol{Y}$.

$$\boldsymbol{Y} = DRT(X_p) \tag{3}$$

In this study, a serial of methods including graph-based, decompositions, clustering, and embedding are used to generate the low dimensional representation of linear and nonlinear data $\boldsymbol{Y}$. This can improve the overall performance of the predictive analytics. Reduced feature sets belonging to different modalities are used as input to ML models (classifier) $C$ for disease prediction and classification tasks (Figure 2).

$$Results = C(Y) \tag{4}$$

A multi-source data fusion framework can collect and integrate data and information collected from multiple sources $S_i$

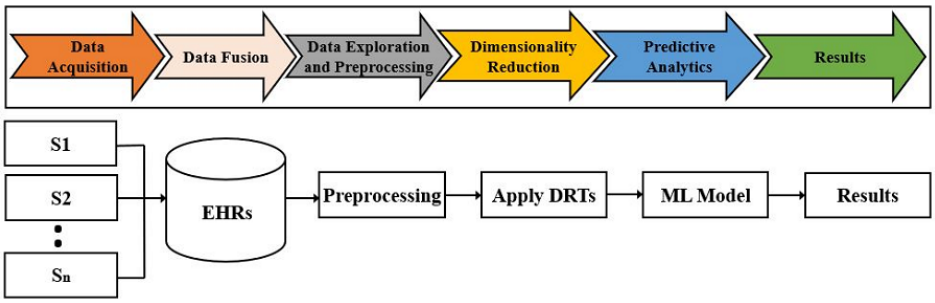


**FIGURE 2.** Proposed paradigm of multi-source data fusion framework. First, data $x_i$ from different sources $S_i$ are combined and stored in EHR. Then, data $X$ is processed using different preprocessing techniques $P$. Next, DRT is applied on preprocessed data $Z$. At the end, predictive analytics model is applied on data $Y$ to attain the results.

to understand the phenomena of interest. For example, data collected from multiple sources, such as patient's demographic information, lab test reports, and the medical images (e.g., X-rays, CT scans, ECG reports), and signals data (e.g., ECG recording, cough or breathing, and other real-time monitoring systems) can help for the prediction of disease. Instead of making decisions manually for each patient, the proposed multi-source fusion framework can fuse (combine) and process data $\boldsymbol{X}$ from multiple sources $S_i$ for efficient disease prediction systems. These systems can assist the physicians for efficient decision making for the medical recommendations.

The main problem with this framework is the diversity of data that needs different strategies for preprocessing, dimensionality reduction, and ML models for the efficient utilization of the most relevant features belonging to multiple sources $S_i$ and modalities $\boldsymbol{M}$.

### B. MULTI-MODAL FEATURE FUSION FRAMEWORK

Uni-modal approaches can only process one type of modality at a time, which restricts their use in the presence of multi-modal data. Multi-source fusion approaches offer an opportunity to work with data belonging to different sources $S_i$. It is difficult to apply any specific data preprocessing mechanism to data collected from multiple sources $S_i$ belonging to various modalities $\boldsymbol{M}$ and data types. To overcome the limitations of the multi-source fusion framework (Figure 2), we proposed a multi-modal feature fusion framework to improve predictive analytics (Figure 3).

Feature level fusion of multiple modalities $\boldsymbol{M}$ is a challenging task, as each modality $m_j$ can have various types of issues like missing values, high dimensionality, sparse datasets, data redundancy, noise, missing or lost signals, quality of signals, lingual issues, lexical and semantic problem, class imbalance etc. To overcome these issues, separate preprocessing $p_j$ of data belonging to each modality $m_j$, and data type is essential before applying any technique for analysis. The proposed multi-modal data and feature fusion framework comprises of the standard data processing pipeline including the acquisition of data from multiple sources $S_i$ and modalities $m_j$, and categorize them as tabular/structured, signals, medical images, and clinical notes (Figure 3).

$$\boldsymbol{X} = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{i,j} \tag{5}$$

where $x_i \in S_i$ derived from source $S_i$ having modality $m_j$.

At the preprocessing phase $p_j$, different preprocessing and data exploration strategies are applied for each data type belonging to multiple modalities and sources separately to normalize and standardize data for further processing.

$$X_{p,j} \geq X_j \tag{6}$$

Then relevant features belonging to various types of data and information collected from multiple sources and modalities are fused in common space at the feature fusion phase to understand the phenomena of interest.

$$\boldsymbol{X}_f = \sum_{m=1}^{m} x_{pj} \tag{7}$$

Next, DRT is applied to obtain reduced dimensions of data $X_{p,j}$.

$$Y_{p,j} = DRT(X_{p,j}) \tag{8}$$

The reduced feature set(s) $Y_{p,j}$ are forwarded to the predictive analytics phase, where ML models are applied for disease prediction and classification tasks. Finally, classifier $C$ is employed on reduced dataset to attain results.

$$Results = C(Y_{p,j}) \tag{9}$$

The use of suitable DRTs can improve the performance of the ML models. Standard evaluation techniques can be applied to evaluate the results.

### C. KNOWLEDGE FUSION FRAMEWORK

The downside of feature fusion is that all the features fused in a common space may not be further processed with common DRTs and ML models to enhance the performance. Moreover, the fundamental challenge in the fusion (combining) of disparate modalities lies in reducing the dimensionality of data and exploring potential features. Simple methods just concatenate the feature vectors to fuse multiple images.

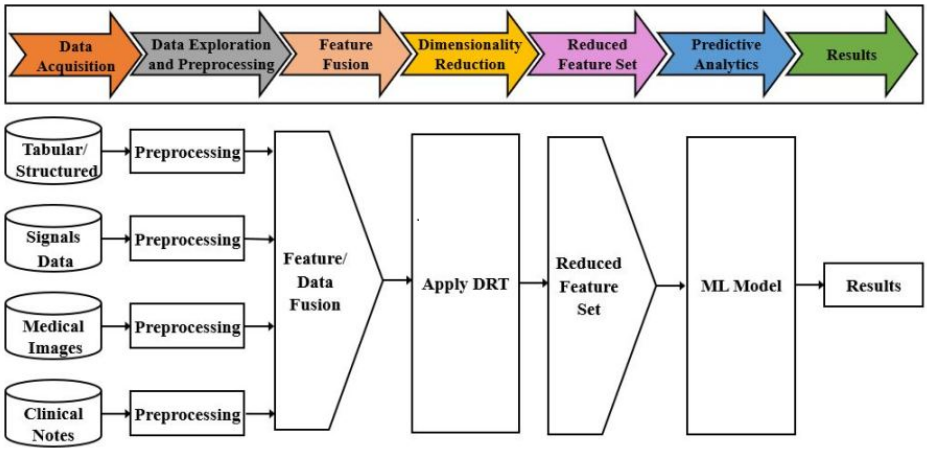


**FIGURE 3.** Structure of proposed multi-modal data fusion framework. In this paradigm, data $x_{i,j}$ from different sources $S_i$ and modalities $m_j$ are processed using different preprocessing techniques $p_i$ according to type of data. Then, DRT is employed to fused preprocessed data to acquire reduced feature set $Y$. The results are obtained after applying appropriate predictive analytics approach.

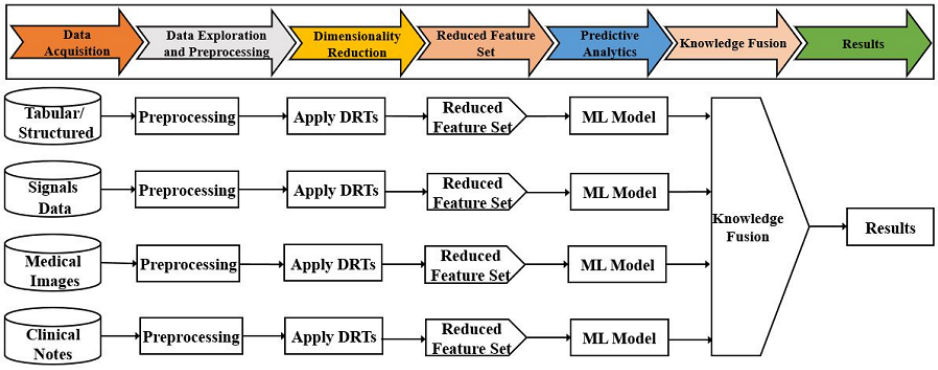


**FIGURE 4.** Architecture of knowledge fusion framework. The proposed framework transforms the data $x_{i,j}$ into reduced feature set $y_{i,j}$ after preprocessing. Then, different machine learning models are employed for each reduced feature set $y_{i,j}$. The results of these models are combined to obtain knowledge fusion.

However, these methods do not consider the varying data types (e.g., image, text, signals) which need further consideration to improve the quality of multi-source and multi-nature data. In this regard, a knowledge fusion framework has been proposed in this study (Figure 4).

The proposed knowledge fusion framework comprises various phases of the standard data processing pipeline (Figure 4). It starts from acquiring data from multiple sources $S$ and modalities $M$ and categorizes them according to data types, such as tabular/structured, signals, medical images, and clinical notes. At preprocessing phase $P$, different preprocessing and data exploration strategies are applied (as shown in Figure 3) for each data type belonging to multiple modalities and sources separately to normalize and standardize data before further processing. Instead of using common feature space for features $F$ belonging to different modalities $M$ and data types, separate steps are performed for each modality $m_j$.

$$X_{p,j} \geq X_{i,j} \tag{10}$$

To reduce the dimensionality of data and extract relevant features $F$ belonging to various types of data and information collected from multiple sources $S$ and modalities $M$, a collection of DRTs according to the type and nature of data was applied for each data type separately.

$$X_p \geq X \tag{11}$$

Next, DRT is applied to obtain reduced data $Y$.

$$Y_{p,j} = DRT(X_{p,j}) \tag{12}$$

DRTs transform the input data $\boldsymbol{X}$ to a low-dimensional representation $\boldsymbol{Y}$ while preserving the original context of information and pairwise relationships between data points. Typically, these pairwise relationships can be preserved via similarity measure or distance-based approaches when combining data from multiple datasets $x_i$ [30]. However, different calculations such as variance and covariance are considered for ICA, PCA, and other DRTs depending on their functionality [11]. Similarly, different DRTs, according to the type and nature of data are selected to enhance the performance of ML models used for predictive analytics. Then reduced and relevant feature sets $y_{i,j}$ belonging to each data type from multiple sources $S$ and modalities $M$ are generated.

The reduced feature set(s) $y_{i,j}$ belonging to each data type is forwarded to the predictive analytics phase. ML models are applied for disease prediction and classification tasks for each data type separately at this phase.

$$Results = C_w(Y_{p,j}) \tag{13}$$

Finally, the results of different classifiers are analyzed as individual classifier and a combination of classifiers can be fused at knowledge fusion phase.

$$K_f Results = C_w(Y_{p,j}) \tag{14}$$

Most of the disease prediction models are constructed using a single ML model. However, the selection of an appropriate ML model is often problematic [63]. Furthermore, the absence of a standardized evaluation techniques for the performance of the classifier also complicates the decision-making process. In addition, multi-modal features, when used as input, may lead to different combinations between features and ML models. This has further increased the complexity of the problem. Different ML models may decode different information. One should select a more reliable model by maximizing the utilization of attained information rather than selecting an optimal one from the available classifiers.

A suitable combination of DRTs and ML models improves the accuracy of results generated using different ML models. Results generated by different ML models for different data types are combined in the knowledge fusion phase, which improves the reliability of knowledge gained for predictive analytics and decision-making. The proposed knowledge fusion framework combines the decision of multiple ML models and utilizes this information for effective decision-making. However, the fusion (mixing) of multi-source, multi-nature, and multi-modal data having high dimensionality is also a big challenge. It is dependent on approaches used for fusion (combining data). Despite all the challenges, the knowledge fusion framework can offer better results when compared with uni-modal and multi-source fusion frameworks to process HDD found in the health informatics domain.

### D. REDUCED FEATURE SETS FUSION FRAMEWORK

For the efficient utilization of multi-source, multi-modal, and multi-nature data $\boldsymbol{X}$ found in different data types, a reduced feature sets fusion framework is proposed (Figure 5). It comprises phases of standard data processing pipeline starting with the acquisition of data from multiple sources $S$ and modalities $M$. This framework also applies separate data preprocessing and exploration operations for each data type at preprocessing phase $P$ similar to the multi-modal feature and knowledge fusion frameworks. After preprocessing and exploring data and features belonging to different modalities and data types, there can be multiple possibilities for further processing data to obtain the reduced representation of multi-nature data $Y$.

$$X = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{i,j} \tag{15}$$

For each $x_{i,j}$, preprocessing was performed where

$$x_{p,j} \geq x_j \tag{16}$$

$$x_f = \sum_{x_p,j} \tag{17}$$

$$X_p \geq X \tag{18}$$

There are two possibilities either apply DRTs or combine the features (fusion) according to the problem statement.

DRT is applied to preprocessed features $p$ belong to $s_i \in m_j$ to obtain reduced dimensions of data $Y$ as $h_u$.

$$Y = DRT(X_p) \tag{19}$$

$$Fusion\ \boldsymbol{F} = X_{p,j} \tag{20}$$

Apply DRTs on fused feature set $\boldsymbol{F}$

$$Y' = DRT(\boldsymbol{F}) \tag{21}$$

Multiple classifier $Cw$ where $w = 1tob$ can be employed on different reduced datasets belonging to different modalities to attain results. The results of different classifiers can be combined using majority voting method.

$$Results = Cw(Y') \tag{22}$$

Features in different modalities can be combined in different ways. One possibility is to combine the features of multiple modalities found in different data types (e.g., tabular and signals data) (Figure 5). This representation is similar to the multi-modal feature fusion framework (Figure 3). Then apply DRT or a combination of DRTs to reduce the dimensionality of data and fuse it in the reduced feature set fusion phase. Another way is to apply different DRTs to reduce the dimensionality of data having different data types (e.g., medical images and clinical notes) as shown in Figure 5 and then fuse it in the reduced feature set fusion phase. Feature sets reduced in both steps are fused (combined) in the reduced feature set fusion phase and forwarded to the predictive analytics phase for prediction and classification tasks. For instance, if the feature sets of modalities cannot be combined before applying DRTs, then the same DRT cannot be applied to the specific data type. According to the problem under consideration, it may require different DRTs for different modalities and data types to reduce the dimensionality of the selected data or combine two or more modalities for DRTs.

Moreover, this framework offers better representations of features in lower dimensions if different DRTs are applied separately to each type of data or combine two or more modalities, according to the problem under consideration. The proposed reduced feature set fusion framework perform the preprocessing for all modalities separately to achieve this objective. Then the features of two or more modalities having different data types (e.g., image and text data) can be combined before applying DRTs.

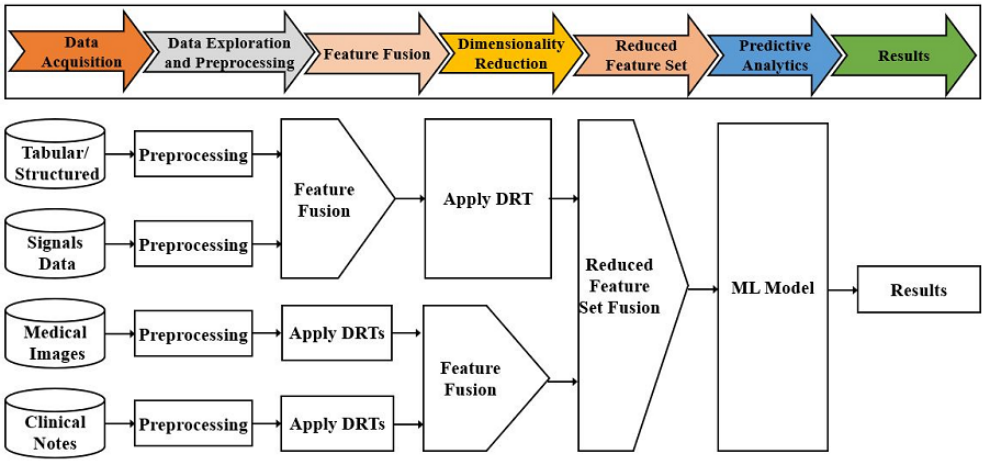


**FIGURE 5.** Proposed reduced feature set fusion framework. Data $x_{i,j}$ can be preprocessed using different preprocessing techniques $p_i$ depending on the modality of data. Reduced features set can be obtained and fused before or after applying DRTs depending on modality $m_j$ of data. Finally, machine learning model is employed to obtain the results.

The proposed reduced feature set fusion framework offers an opportunity to fuse data before or after applying DRT and combine the reduced feature set to explore the best results. It is considered that the ML model applied on the optimal (reduced) feature set can improve the performance of ML models. This fusion framework needs extra attention for exploring data. Data comprises multiple modalities, sources, or data types can be combined before and after applying DRTs to get the optimal feature sets for analysis and predictive analytics. Moreover, a common ML model applied on multi-source, multi-modal, and different data types and reduced representations may not improve the performance of the specific ML model. Despite all these improvements and flexibility to develop reduced feature sets and their fusion, this framework faces challenges for the fusion of multiple reduced features sets, their representations, and the selection of a common ML model for different reduced feature sets.

### E. DECISION FUSION FRAMEWORK BASED ON A HYBRID APPROACH

A decision fusion framework based on a hybrid fusion approach is proposed in this study to overcome the issues and challenges found for the fusion of multi-source, multi-modal, and different data types of multi-nature health informatics data (Figure 6). Hybrid fusions increase the flexibility by offering an opportunity to dynamically applying DRTs on different data types, combine features of multi-source, multi-modal, HDD before and after applying DRTs. Reduced feature sets are forwarded to the predictive analytical phase, where most efficient ML models are applied to reduced feature sets of different data types to explore effective decisions.

Decisions generated via different ML models are fused to make a collective decision (in some situations, it may look like the knowledge fusion framework (Figure 4). However, multiple combinations of data $\boldsymbol{X}$ and features $\boldsymbol{F}$ belonging to multi-source, multi-modal, HDD, and multi-nature require a mechanism for selecting the most suitable and compatible techniques to improve the accuracy of decisions. The fusion of decisions generated by multiple ML models needs human interaction to get a common decision to improve the accuracy of decisions for disease prediction and management. Multiple ML models are typically used when combining data from multiple sources $S$ to get a uniform representation in a common decesion space. The proposed decision fusion framework can represent the results generated by different ML models (Figure 6).

Next, DRT is applied to obtain reduced data $Y$. There are two possibilities either apply DRTs or combine the features (fusion) according to the problem statement. DRT is applied to preprocessed features $p$ belong to $s_i \in m_j$ to obtain reduced dimensions of data $Y$ as $h_u$. Apply DRTs on fused feature set $\boldsymbol{F}$ Individual or multiple classifiers $CorCw$ can be employed on different reduced datasets belonging to different modalities to attain results. The results of different classifiers can be combined using majority voting method.

The results generated by different ML models using fused features or independent data sources are fused (combined) to improve the accuracy and reliability of decisions. The decision fusion framework offers an opportunity to diagnose disease based on the preserved information to detect new patients with similar symptoms in less time. For this purpose, computation of probabilities decides how different imaging (medical imaging) and non-imaging (patient demographic information and diagnostic history) modalities $M$ can be used for decision fusion to improve the performance.

However, selecting the right fusion level depends on fusion strategy and data complexity to achieve the required results. A mechanism is needed to ensure dynamic updates when new data arrives for analysis. DRTs works well to eliminate the redundant and non-significant features before and after the fusion of multi-modal, multi-source, multi-nature data.

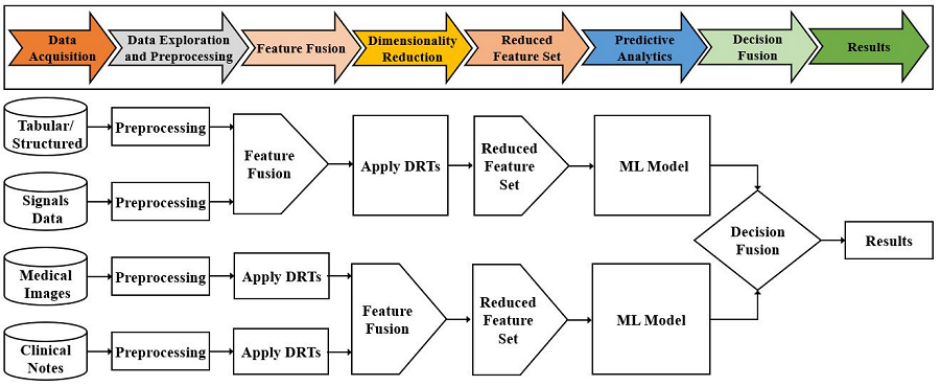


**FIGURE 6.** Structure of proposed decision fusion framework based on hybrid approach. Preprocessed data $x_{i,j}$ for each modality $m_j$ can be fused and transformed into reduced feature set before or after applying DRT. Then, decision of different predictive analytics models can be used to obtain decision fusion.

## IV. EXPERIMENTAL EVALUATIONS OF THE FUSION FRAMEWORKS AND DISCUSSION

In this study, we explore the efficiency of the proposed fusion frameworks for the detection of COVID-19 cases. Experimental work was performed using COVID19 Chest X-ray dataset [64]. The dataset holds 178 normal and 95 COVID-19 cases. The dataset comprises three modalities, e.g., X-ray images, patient information as tabular data, and clinical notes, which are helpful for the prediction of COVID cases. Dataset consist of tabular, image, and text data which hold different information for the detection of COVID-19 cases. To explore the significance of each modality (data type) for disease detection different combinations of these modalities were used according to the proposed fusion frameworks.

During the preprocessing phase $P$, discrepancies from data $X$ are removed for each modality $m_j$ to ensure the completeness of data. This step is necessary before applying any analysis technique and achieving the optimal outcomes (results). Moreover, it improves the efficiency of ML models used for classification and prediction. In this phase, we applied a filter to identify the missing values and use the majority vote method [65], to impute the missing values. To overcome the imbalanced class issue, we eliminated the classes of data having five or fewer samples. The diversity of data caused due to multiple labeling were identified and converted into a uniform format, such as categorical data is converted into numeric data to prepare it for analysis. Then we split the whole dataset into train and test sets.

Figure 7 represent the significance of features. Different features are ranked according to their significance for disease detection. Highly correlated and least significant features have been eliminated using DRTs.

A collection of DRTs like PCA, SVD, LSA, PP, ICA, LPP, LDA, KPCA, LLE, SOM, LVQ, and t-SNE are applied to reduce the dimensionality of the data. It also helps to explore the best suitable DRTs according to the nature and type of data. Different classifiers and ML models are trained and tested on reduced feature sets. The prediction accuracy of the ML model indicate the performance accuracy and the best combination of DRTs and ML models for prediction and classification task. Moreover, different ML models were trained and tested to develop an efficient disease prediction model. The compatibility of DRTs and ML models should be kept in view when selecting a combination of DRTs and ML models to achieve the highest performance accuracy. Accuracy of the models is calculated as number of right predictions divided by total predictions.

Table 2 presents the result where collection of DRTs were used to reduce the dimensionality of various data types and reduced feature set were sent to KNN (ML model). PCA achieved highest accuracy score 97% for Tab data. For Img data Isomap achieved accuracy score of 96% and LSA achieved 95% accuracy for Txt data. Similarly, bold value indicate highest accuracy for different fusion levels and modalities.These results are obtained for normalized dataset. However, the results of the DRTs and performance of of ML model can change according to the problem statement (e.g., number of reduced dimensions used for analysis) or type and nature of data.

For different fusion levels, modalities, and DRTs SVM attain different accuracies have been shown in Table 3. In unimodal for tab data SVM with PCA achieved highest accuracy of 94%. Isomap achieved highest accuracy of 96% for img data, and for txt data LSA achieved an accuracy of 92%. Similarly, accuracies of different modalities and fusion level after applying DRTs for SVM have been shown. Bold values indicate the highest accuracy.

Table 4 represents different accuracies CNN model achieved after applying a collection DRTs for different modalities. In unimodal CNN for tab data with PCA achieved highest accuracy of 95%. t-SNE d highest accuracy of 97% for img data, and for txt data LSA achieved an accuracy of 95%. Accuracies attained for different modalities and fusion level after applying DRTs with CNN have been shown (Table 4). Bold values represent the highest accuracy value.

After applying a collection of DRTs, Table 5 indicates the different accuracies of RF with different modalities and fusion levels. RF with PCA attained the highest accuracy of

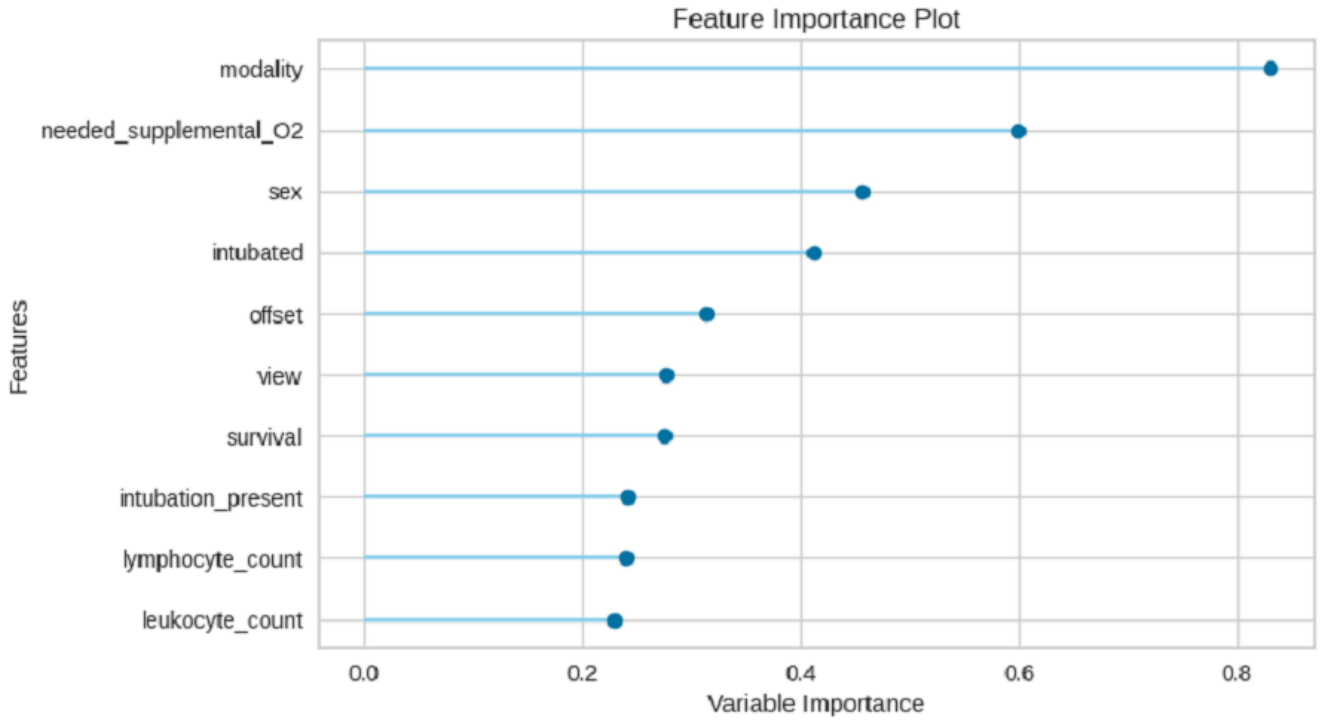


**FIGURE 7.** Data processing (Feature Importance).

94% for tab data. t-SNE attained the highest accuracy of 97% for img data, and LSA attained an accuracy of 95% for txt data. In the same way, accuracies for different modalities and fusion level after applying DRTs with RF have been shown. Bold values indicate the highest accuracy.

The performance accuracy of DT has been presented in Table 6 which shows the different accuracy levels of DRT and ML models for the specified fusion level. For tab data DT achieved highest accuracy of 94% with PCA. Isomap achieved highest accuracy of 97% for img data, and for txt data LSA achieved an accuracy of 95%. Similarly, accuracies for different modalities and fusion level after applying DRTs with DT have been shown. Bold values shows the highest accuracy.

Another purpose of these tables is to present different combinations of DRTs and ML models for the proposed fusion frameworks. Similarly, accuracies of different modalities and fusion level after applying DRTs for the selected model helps in selection of best combination.

In knowledge fusion framework, a combination of multiple ML models were analyzed for fusion framework including SVM, RF, CNN, KNN, and DT. Grid search and cross validation techniques were applied to tune the hyper parameters for DRTs and ML models. This combination helps to explore the actionable insights from the HDD, multi-nature, multi-modal, and multi-source health informatics data for efficient predictive analytics (disease prediction and management systems).

For the decision fusion, best results of different ML models and DRTs are combined to ensure the accuracy of decisions for disease prediction and management. In this study, we follow major voting method [66] to combine the results of different ML models(KNN, SVM, CNN, RF, DT). Parameters of individual ML model were decided using grid search approach. To best of our knowledge, only few studies proposed decision fusion framework for disease prediction. In this study we proposed a decision fusion framework to enhance the accuracy of prediction system. According to the study [67], structure of different classifiers may generate ambivalent results. Due to the limitations of conceptual framework and Dieterich's reasons (statistical, computational, and representational) which indicate the need of novel mechanisms to improve the accuracy of the disease prediction system [66]. In this study, we justify how multiple classifiers may offer better decision than a single one classifier.

With the aim to improve the decision for disease prediction we proposed a decision fusion framework. To evaluate the advantage of using multiple classifiers for decision fusion versus a single classifier approach for disease detection. The proposed framework is a justification of concept that for multi-source and multi-modal HDD no single classifier can be suggested. As different classifiers generate varying results and have different capabilities which can be best utilized via decision fusion to enhance the performance and reliability of disease prediction systems. For this purpose, we attain result for both scenarios apply 5 classifiers (KNN, SVM, CNN, DT, and RF) for different modalities and apply a collection of DRTs and fuse the results of multiple classifiers (KNN, SVM, CNN, DT, and RF) via majority voting method. This fused

**TABLE 2.** Accuracy score of KNN with multiple DRTs in different fusion frameworks.

| Type of Fusion | Modality used | PCA | SVD | ICA | LDA | LSA | PP | LPP | KPCA | Isomap | MDS | LLE | SOM | LVQ | t-SNE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unimodal | Tab | **.97** | .79 | .81 | .69 | .78 | .67 | .60 | .89 | .87 | .60 | .73 | .88 | .67 | .86 |
| | Img | .94 | .75 | 79 | .63 | .69 | .57 | .62 | .83 | **.96** | .74 | .93 | .86 | .75 | .90 |
| | Txt | .83 | .81 | .62 | .70 | **.92** | .64 | .81 | .83 | .84 | .63 | .67 | .64 | .60 | .82 |
| Multi-source fusion | Tab+Img | **.96** | .87 | .76 | .70 | .59 | .71 | .69 | .81 | .83 | .78 | .77 | .81 | .69 | .94 |
| | Tab+Txt | 88 | .75 | 66 | .73 | .86 | .57 | .62 | .81 | **.93** | .82 | .73 | .76 | .85 | .87 |
| | Txt+Img | .85 | .84 | .63 | **.94** | .63 | .62 | .61 | .83 | .64 | .73 | .86 | .90 | .87 | .84 |
| | Tab+Img+Txt | **.95** | .84 | .78 | .79 | .67 | .68 | .71 | .86 | .81 | .73 | .76 | .78 | .67 | .90 |
| Multi-modal fusion | Tab+Img | **.94** | .59 | .81 | .69 | .78 | .78 | .66 | .87 | .64 | .90 | .93 | .91 | .92 | .90 |
| | Tab+Txt | .89 | .75 | 59 | .63 | .80 | .57 | .62 | .81 | .82 | **.92** | .83 | .86 | .85 | .89 |
| | Txt+Img | .91 | .70 | .64 | .73 | .86 | .90 | .65 | .64 | **.94** | .90 | .63 | .62 | .87 | .84 |
| | Tab+Img+Txt | .94 | .65 | .64 | .63 | .94 | .63 | .93 | **.96** | .95 | .82 | .61 | .90 | .57 | .84 |
| Knowledge fusion | Tab+Img | **.95** | .89 | .87 | .69 | .78 | .85 | .84 | .63 | .80 | .83 | .81 | .78 | 67 | .91 |
| | Tab+Txt | **.91** | .87 | 67 | .63 | .69 | .57 | .62 | .89 | .86 | .79 | .83 | .87 | .65 | .87 |
| | Txt+Img | .93 | .64 | .63 | **.94** | .63 | .62 | .61 | .87 | .84 | .73 | .86 | .90 | .87 | .81 |
| | Tab+Img+Txt | **.96** | .78 | .63 | .64 | .63 | .58 | .71 | .88 | .74 | .83 | .76 | .81 | .66 | .84 |
| Reduced feature set fusion | Tab+Img | .87 | .79 | .61 | .69 | .78 | 76 | .68 | .59 | .64 | **.90** | .63 | .71 | .62 | .80 |
| | Tab+Txt | .84 | .75 | 69 | .63 | .69 | **.87** | .62 | .71 | .64 | .63 | .74 | .63 | .62 | .80 |
| | Txt+Img | .83 | .71 | .60 | .64 | .73 | .86 | .80 | .77 | .83 | .64 | .63 | .58 | .83 | **.89** |
| | Tab+Img+Txt | .86 | .64 | .63 | .76 | .63 | .62 | .61 | .83 | **.91** | .86 | .73 | .87 | .65 | .89 |

**TABLE 3.** Accuracy score of SVM with multiple DRTs in different fusion frameworks.

| Type of Fusion | Modality used | PCA | SVD | ICA | LDA | LSA | PP | LPP | KPCA | Isomap | MDS | LLE | SOM | LVQ | t-SNE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unimodal | Tab | **.94** | .59 | .81 | .69 | .78 | .76 | .79 | .84 | .78 | .80 | .83 | .77 | .62 | .80 |
| | Img | .92 | .81 | 69 | .83 | .69 | .57 | .62 | .71 | .95 | .94 | .93 | **.96** | .75 | .89 |
| | Txt | .83 | .90 | .63 | .84 | **.92** | .62 | .71 | .80 | .84 | .83 | .86 | .70 | .57 | .84 |
| Multi-source fusion | Tab+Img | **.96** | .79 | .81 | .79 | .78 | .68 | .69 | .87 | .65 | .79 | .83 | .71 | .52 | .90 |
| | Tab+Txt | .89 | .75 | 69 | .33 | .76 | .57 | .75 | .72 | .82 | .74 | .83 | .86 | .75 | **.90** |
| | Txt+Img | .91 | .84 | .73 | .67 | .75 | .62 | .71 | .67 | **.94** | .80 | .81 | .89 | .77 | .83 |
| | Tab+Img+Txt | **.97** | .74 | .76 | .89 | .73 | .82 | .71 | .87 | .89 | .83 | .86 | .75 | .71 | .84 |
| Multi-modal fusion | Tab+Img | **.93** | .86 | .81 | .78 | .76 | .88 | .78 | .83 | .81 | .70 | .83 | .82 | .62 | .91 |
| | Tab+Txt | .89 | .86 | .88 | .69 | .79 | .87 | .72 | .81 | **.92** | .84 | .83 | .86 | .65 | .89 |
| | Txt+Img | **.90** | .84 | .75 | .84 | .76 | .82 | .77 | .74 | .84 | .73 | .86 | .87 | .77 | .84 |
| | Tab+Img+Txt | **.95** | .84 | .73 | .84 | .75 | .78 | .83 | .87 | .74 | .81 | .82 | .85 | .67 | .89 |
| Knowledge fusion | Tab+Img | .91 | .89 | .82 | .78 | .81 | .77 | .78 | .82 | **.92** | .89 | .83 | .71 | .62 | .87 |
| | Tab+Txt | .88 | .85 | 79 | .83 | .89 | .78 | .67 | .79 | .87 | .74 | .83 | .76 | .78 | **.91** |
| | Txt+Img | .85 | .81 | .83 | .84 | .73 | .82 | .75 | .70 | .74 | .73 | .81 | **.88** | .87 | .84 |
| | Tab+Img+Txt | **.94** | .86 | .83 | .79 | .81 | .80 | .84 | .80 | .78 | .83 | .96 | .75 | .84 | .93 |
| Reduced feature set fusion | Tab+Img | .86 | .59 | .81 | .79 | .76 | 75 | .67 | .79 | .68 | .79 | .83 | .81 | .72 | **.90** |
| | Tab+Txt | **.88** | .76 | 79 | .73 | .69 | .67 | .62 | .71 | .82 | .84 | .87 | .76 | .65 | .86 |
| | Txt+Img | .85 | .64 | .63 | .94 | .63 | .62 | .61 | .70 | .64 | .73 | .86 | **.90** | .87 | .84 |
| | Tab+Img+Txt | **.91** | .84 | .83 | .74 | .65 | .66 | .71 | .70 | .74 | .73 | .87 | .85 | .77 | .88 |

**TABLE 4.** Accuracy score of CNN with multiple DRTs in different fusion frameworks.

| Type of Fusion | Modality used | PCA | SVD | ICA | LDA | LSA | PP | LPP | KPCA | Isomap | MDS | LLE | SOM | LVQ | t-SNE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unimodal | Tab | **.95** | .79 | .81 | .89 | .87 | .78 | .85 | .74 | .83 | .89 | .82 | .89 | .82 | .90 |
| | Img | .91 | .85 | 69 | .83 | .69 | .87 | .82 | .71 | .82 | .74 | .86 | .75 | .93 | **.97** |
| | Txt | .92 | .64 | .83 | .84 | **.95** | .76 | .61 | .70 | .64 | .83 | .86 | .84 | .82 | .89 |
| Multi-source fusion | Tab+Img | .90 | .79 | .88 | .79 | .78 | .78 | .79 | .87 | .64 | .90 | **.93** | .91 | .92 | .90 |
| | Tab+Txt | .87 | .75 | .84 | .73 | **.90** | .87 | .62 | .81 | .82 | .79 | .83 | .86 | .75 | .89 |
| | Txt+Img | .92 | .64 | .83 | .89 | .82 | .86 | .81 | .87 | **.94** | .83 | .86 | .89 | .71 | .84 |
| | Tab+Img+Txt | **.95** | .84 | .73 | .84 | .73 | .68 | .71 | .93 | .91 | .83 | .78 | .95 | .67 | .93 |
| Multi-modal fusion | Tab+Img | **.94** | .87 | .81 | .69 | .81 | .78 | .86 | .81 | .84 | .78 | .83 | .81 | .88 | .90 |
| | Tab+Txt | .90 | .86 | 69 | .83 | **.91** | .87 | .62 | .71 | .92 | .84 | .73 | .86 | .75 | .87 |
| | Txt+Img | .82 | .64 | .73 | .84 | .73 | .86 | .76 | .70 | **.89** | .73 | .86 | .70 | .67 | .86 |
| | Tab+Img+Txt | **.96** | .84 | .83 | .81 | .77 | .82 | .80 | .78 | .89 | .83 | .86 | .85 | .67 | .84 |
| Knowledge fusion | Tab+Img | **.93** | .89 | .81 | .83 | .88 | .87 | .78 | .69 | .86 | .80 | .83 | .81 | .82 | .90 |
| | Tab+Txt | .91 | .85 | 89 | .78 | .89 | .78 | .82 | .71 | .73 | .84 | .83 | .86 | .66 | **.92** |
| | Txt+Img | .91 | .84 | .83 | .79 | .83 | .76 | .81 | .78 | **.94** | .81 | .85 | .83 | .87 | .89 |
| | Tab+Img+Txt | **.96** | .86 | .81 | .78 | .81 | .68 | .77 | .87 | .94 | .93 | .90 | .85 | .77 | .91 |
| Reduced feature set fusion | Tab+Img | .86 | .79 | .81 | .69 | .78 | 77 | .78 | .59 | .84 | .80 | .76 | .81 | .72 | **.89** |
| | Tab+Txt | .89 | .75 | 69 | .63 | .69 | .57 | .62 | .81 | .72 | .74 | **.93** | .76 | .65 | .79 |
| | Txt+Img | .92 | .64 | .63 | .94 | .63 | .62 | .61 | .70 | .64 | **.93** | .86 | .70 | .77 | .74 |
| | Tab+Img+Txt | .91 | .76 | .71 | .94 | .73 | .64 | .67 | .70 | **.94** | .66 | .62 | .58 | .63 | .70 |

decision is more accurate and reliable for disease prediction. Moreover, the results based on reduced feature set attained after applying DRTs. In this study, the prediction accuracy of the model is used to weight the vote. The accuracy of the ML model was calculated using sensitivity, specificity, the Area Under Curve (AUC), and Receiver Operating Characteristics (ROC) [68]. To weight the vote, the participation of decision got increase with high sensitivity and specificity score.

The results of SVM, CNN, KNN, DT, and RF are compared and fused for decision fusion. For instance, highest accuracy of different individual classifiers for Tab data are as follow KNN with PCA is 97%, SVM with PCA is 94%, CNN with PCA is 95%, RF with LPP is 90%, and DT with PCA is 92%, respectively. For Img data accuracy of individual classifier KNN with Isomap is 96%, SVM with SOM is 96%, CNN with t-SNE is 97%, RF with PCA is 91%, and DT with

**TABLE 5.** Accuracy score of RF with multiple DRTs in different fusion frameworks.

| Type of Fusion | Modality used | PCA | SVD | ICA | LDA | LSA | PP | LPP | KPCA | Isomap | MDS | LLE | SOM | LVQ | t-SNE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unimodal | Tab | .89 | .79 | .86 | .87 | .78 | .88 | **.90** | .84 | .87 | .83 | .76 | .78 | .82 | .87 |
| | Img | **.91** | .85 | 79 | .63 | .69 | .77 | .86 | .79 | .87 | .81 | .83 | .86 | .83 | .85 |
| | Txt | .84 | .76 | .77 | .84 | .63 | .72 | .76 | .78 | **.88** | .72 | .76 | .79 | .66 | .86 |
| Multi-source fusion | Tab+Img | .90 | .87 | .81 | .79 | .83 | .88 | .79 | **.92** | .64 | .78 | .83 | .81 | .80 | .87 |
| | Tab+Txt | .93 | .75 | 69 | .63 | .69 | .87 | .62 | .71 | .92 | .94 | .93 | **.96** | .85 | .89 |
| | Txt+Img | .91 | .86 | .83 | .44 | .83 | .88 | .71 | .70 | .84 | **.93** | .86 | .90 | .77 | .90 |
| | Tab+Img+Txt | .92 | .64 | .63 | .94 | .63 | .62 | .61 | .70 | **.94** | .93 | .89 | .85 | .87 | .92 |
| Multi-modal fusion | Tab+Img | 90 | .79 | .81 | .69 | .78 | .78 | .66 | .87 | .64 | .90 | .83 | .81 | .62 | **.92** |
| | Tab+Txt | .90 | .85 | 89 | .73 | .86 | .77 | .62 | .71 | **.92** | .84 | .79 | .86 | .65 | .91 |
| | Txt+Img | .65 | .64 | .63 | **.94** | .63 | .62 | .61 | .70 | .64 | .83 | .86 | .90 | .87 | .89 |
| | Tab+Img+Txt | .90 | .64 | .63 | .89 | .63 | .62 | .61 | .70 | **.94** | .93 | .88 | .85 | .77 | .84 |
| Knowledge fusion | Tab+Img | .76 | .59 | .81 | .69 | .78 | .87 | .78 | .69 | .64 | .90 | **.93** | .91 | .82 | .90 |
| | Tab+Txt | .88 | .75 | 69 | .63 | .69 | .57 | .62 | .71 | **.91** | .94 | .93 | .96 | .95 | .89 |
| | Txt+Img | .85 | .64 | .63 | .84 | .63 | .62 | .61 | .70 | .64 | **.93** | .86 | .90 | .87 | .84 |
| | Tab+Img+Txt | .89 | .64 | .63 | .94 | .63 | .62 | .61 | .70 | .89 | **.93** | .92 | .85 | .77 | .84 |
| Reduced feature set fusion | Tab+Img | .90 | .59 | .81 | .69 | .78 | 86 | .78 | .59 | .64 | .90 | **.93** | .91 | .76 | .90 |
| | Tab+Txt | .89 | .75 | 69 | .63 | .79 | .57 | .62 | .71 | .92 | **.94** | .91 | .86 | .67 | .89 |
| | Txt+Img | .92 | .64 | .63 | .94 | .63 | .62 | .61 | .70 | .64 | **.93** | .86 | .90 | .87 | .84 |
| | Tab+Img+Txt | .93 | .64 | .63 | .94 | .63 | .62 | .61 | .70 | .89 | .93 | .86 | .88 | .77 | **.94** |

**TABLE 6.** Accuracy score of DT with multiple DRTs in different fusion frameworks.

| Type of Fusion | Modality used | PCA | SVD | ICA | LDA | LSA | PP | LPP | KPCA | Isomap | MDS | LLE | SOM | LVQ | t-SNE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unimodal | Tab | **.92** | .89 | .81 | .79 | .78 | .88 | .79 | .64 | .87 | .90 | .90 | .88 | .62 | .90 |
| | Img | .90 | .75 | 69 | .63 | .69 | .57 | .62 | .71 | **.92** | .84 | .90 | .86 | .65 | .89 |
| | Txt | .85 | .64 | .76 | .84 | .73 | .67 | .61 | .70 | .64 | **.90** | .76 | .70 | .57 | .84 |
| Multi-source fusion | Tab+Img | **.93** | .79 | .81 | .69 | .79 | .78 | .82 | .87 | .84 | .90 | .83 | .91 | .62 | .90 |
| | Tab+Txt | .89 | .75 | 69 | .63 | .69 | .57 | .62 | .71 | .92 | **.94** | .83 | .86 | .65 | .89 |
| | Txt+Img | .91 | .87 | .83 | .89 | .78 | .83 | .81 | .87 | .64 | **.92** | .86 | .90 | .67 | .84 |
| | Tab+Img+Txt | .90 | .64 | .77 | .84 | .82 | .86 | .76 | .70 | .91 | .93 | .89 | .85 | .67 | **.92** |
| Multi-modal fusion | Tab+Img | .86 | .79 | .83 | .85 | .77 | .87 | .86 | .81 | .84 | .89 | .83 | .81 | .62 | **.90** |
| | Tab+Txt | .89 | .75 | 69 | .63 | .69 | .57 | .62 | .87 | .82 | **.91** | .88 | .76 | .57 | .89 |
| | Txt+Img | .80 | .76 | .73 | **.89** | .66 | .62 | .61 | .70 | .64 | .88 | .86 | .80 | .65 | .84 |
| | Tab+Img+Txt | **.90** | .84 | .63 | .74 | .63 | .62 | .61 | .70 | .74 | .78 | .76 | .75 | .50 | .84 |
| Knowledge fusion | Tab+Img | .90 | .79 | .77 | .80 | .78 | .87 | .78 | .69 | **.94** | .90 | .93 | .89 | .62 | .90 |
| | Tab+Txt | .90 | .85 | 89 | .83 | .79 | .87 | .82 | .71 | .81 | .84 | .80 | .86 | .65 | **.92** |
| | Txt+Img | .91 | .93 | .63 | **.94** | .63 | .62 | .61 | .70 | .64 | .73 | .86 | .90 | .87 | .90 |
| | Tab+Img+Txt | **.96** | .64 | .63 | .84 | .68 | .62 | .61 | .70 | .94 | .93 | .76 | .85 | .87 | .94 |
| Reduced feature set fusion | Tab+Img | **.92** | .59 | .81 | .69 | .78 | 86 | .78 | .59 | .64 | .90 | .93 | .91 | .92 | .90 |
| | Tab+Txt | .89 | .75 | 69 | .63 | .69 | .57 | .62 | .71 | .92 | .94 | **.93** | .86 | .75 | .89 |
| | Txt+Img | .90 | .64 | .63 | **.92** | .63 | .62 | .61 | .70 | .64 | .73 | .86 | .90 | .87 | .88 |
| | Tab+Img+Txt | **.90** | .64 | .73 | .83 | .73 | .62 | .61 | .87 | .84 | .93 | .86 | .83 | .66 | .84 |

Isomap is 92%, respectively. For Txt data highest accuracy of individual classifier KNN and SVM with LSA is 92%, CNN with LSA is 95%, RF with Isomap is 88%, and DT with MDS 90%, respectively. According to majority voting method KNN with PCA generate best result for tab data, CNN with t-SNE achieved highest score for image data, and CNN with LSA generate best results. Decision of these classifiers are considered more significant for decision fusion.

The knowledge and decision fusion frameworks are practically useful when data comprises highly heterogeneous, multi-dimensional, and high-dimensional healthcare data. The proposed fusion frameworks will undoubtedly be a milestone for developing dynamic disease prediction and management systems utilizing multi-modal, multi-source, multi-nature, and high-dimensional health informatics data.

## V. OPPORTUNITIES AND CHALLENGES OF DIMENSIONALITY REDUCTION AND FUSION APPROACHES FOR USED PREDICTIVE ANALYTICS IN HEALTH INFORMATICS

Despite all the achievements and successful implementations of DRTs for DR and fusion approaches used for combining multi-modal, multi-sources, high dimensional healthcare data for predictive analytics, these techniques (DRTs and fusion) still face many issues in their adoption and implementation. This section highlights the opportunities and challenges of DRTs and the fusion approaches used for health informatics data. These challenges indicate new opportunities for improvement to research communities.

### A. HEALTHCARE DATA REPRESENTATION AND TRANSFORMATION

Data fusion approaches can select the most relevant features from the most effective modality having an appropriate amount of features for disease prediction. In the real world, most of the healthcare data collected from multiple sources often found in different modalities (format /data types) and has high dimensionality [5], [40]. DRTs can reduce the dimensionality of data and present it in lower dimensions. Many DRTs can work with a specific type of data, e.g., image, text, or signals. Similarly, when fusing data from these diverse kinds of modalities, several issues have been found, such as converting data type, measuring scales, and formats (qualitative and quantitative). In such a situation, fusion and processing of multi-nature data become a challenging task [1], [3]. In some situations, humans can easily perceive such variations in data and make decisions accordingly. While for the same situation, when relying on DRTs and data fusion

tools, it may require many computations and coding even to solve a simple problem. Despite such issues, the development of the ML model that can automatically perform fusion and integration of multi-modal data has become the need of the hour. It will save time for manual interpretation each time new data evolves; furthermore, it also reduces the error rate.

### B. MANAGEMENT OF STREAMING DATA

Another key challenge with data fusion approaches is dealing with multi-nature health-related data [59]. In the healthcare domain, a massive volume of data is originating at a rapid pace every hour. The main benefit of the evolving data is that medical experts can get assistance from automatic disease prediction systems. Data in healthcare collected from different devices and sensors and their integration has become a challenging issue [32]. Fusion approaches should be capable of dynamically combine data belonging to multiple formats and sources [69]. Although, some fusion approaches have been proposed in the literature to manage continuous streaming data. Yet, some challenging and unsettled issues regarding the effective analysis of the large scale and continuously growing data are missing values, blind sources, and noise.

### C. COMPLEXITY AND COMPUTATION

An increase in healthcare data and the diversity found in data has become a big challenge. Similarly, the fusion process is also complex and computationally expensive [3]. To overcome, such issues DRTs offer an effective solution by providing a low-dimensional representation of the data. DRTs also has data interpretation issue after transforming data in lower dimensions. In some situations, using a simple classifier is not enough to meet the requirements of multi-nature data and needs modification. There is a need to improve the ML algorithm to handle complex and varying nature healthcare data.

### D. FUSION STRATEGY

Data fusion approaches are often affected by subsequent phases such as data gathering, preprocessing, and DR. The advantages and disadvantages of fusion may also depend on the strategy used for fusion. Similarly, environmental factors, especially in healthcare, matter a lot [21]. A paper-based approach is in practice in some healthcare organizations, due to which required data may be left from the analysis. Similarly, the selection of suitable fusion strategy according to given requirements is also a big challenge because of varying needs of healthcare data as well as variations in treatments and diagnostic process [5], [31].

### E. HEALTHCARE DATA ANALYTICS

To obtain compact and precise healthcare data for analysis and decision making, selecting reliable data sources with few errors, noise, and missing values is a big challenge. For predictive analytics in healthcare, time-frame, correctness, and reliability of health status measuring devices (e.g., glucometer, ECG device, blood pressure measuring instruments, etc.) and correct recording are essential. Particularly, health informatics data collection, storage, process, and transmission have become challenging and may affect the reliability, consistency, and accuracy of data quality [70].

### F. MEASUREMENT AND EVALUATION OF PERFORMANCE

Fusion of data belonging to multiple sources and modalities measurement and evaluation of performance after applying fusion strategies (feature and knowledge fusion) is another challenging area to consider [59]. More dependency on IoT devices and technology has made performance evaluation more complex. In healthcare, there are many approaches to measure performance based on improved accuracy, reduce computation time, reduce error rate, reduce response time, and significance and reliability of results. Others such as F-measure, Area Under the Curve (AUC), Mean Squared Error (MSE), and Correlation Coefficient (COR) etc [68]. The selection of a suitable measuring approach to improve performance is also a challenging task.

### G. VARIATION IN TERMS AND CONCEPTS

When combining data from different sources, variations often found in terms, abbreviations used, and concepts [71]. Variation in terms such as physician and doctor represent the same entity, while Fly and fly similar words may be used for different purposes (context) such as Fly as a noun and fly as a verb. Such variations are common in health informatics data and need modification according to the concepts and their use. Although Natural Language Processing (NLP) approaches play an influential role in overcoming this issue, it is still a big issue, especially when combining data from different modalities and sources [5].

### H. DIMENSIONALITY REDUCTION AND FUSION OF REDUCED REPRESENTATIONS

When applying DRTs for DR, it may be challenging to answer the key questions: how much dimensionality for the given data set can be reduced? The decision regarding the number of components to use for analysis and decision making. Similarly, how many features or components should be eliminated to avoid or eliminate the noisy data? [72]. Although, most of the time, these questions vary according to the problem statement and nature of the data. It became complicated and time-consuming to evaluate the effects of reduction. Variation in feature selection and extraction methods also affects data representation and aspects used to measure the performance [31]. Moreover, combining the results of different DRTs having different representations, such as PCA represent reduced feature set as PCs. In contrast, ICA as ICs in such situation fusing the reduced representations becomes a challenge.

## VI. CONCLUSION

To combine multi-source, multi-modal healthcare data for predictive analytics, we proposed multi-modal, multi-source, knowledge, and decision fusion frameworks in this study.

We applied a collection of DRTs to reduce the dimensionality of data while preserving the original aspects of HDD. Different ML models were used for disease prediction. The proposed frameworks offer more reliable and authentic results, as it predicts the disease after combining the features of data and the results (outcomes) at different phases. It can be helpful for the development of effective and reliable systems for health informatics. Moreover, fusion approaches with reduced feature sets presented in this work can provide considerable support for developing CDSS, prediction systems, and intelligent health monitoring systems. No doubt, the proposed fusion frameworks offer a clear representation of different fusion levels. The experimental results show that the proposed methodology achieved better accuracy when compared with individual modalities and the state-of-the-art fusion approaches. However, an apt problem for future study is to explore a comprehensive approach to identify the relationships among features influence during data and feature fusion. So, novel approaches can be explored in future.

## REFERENCES

[1] D. Korzun and A. Meigal, "Multi-source data sensing in mobile personalized healthcare systems: Semantic linking and data mining," in *Proc. 24th Conf. Open Innov. Assoc. (FRUCT)*, Apr. 2019, pp. 187–192.
[2] A. K. Gárate-Escamila, A. H. E. Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informat. Med. Unlocked*, vol. 19, Jan. 2020, Art. no. 100330.
[3] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, *arXiv:1806.00064*.
[4] F. Ali, S. El-Sappagh, S. M. R. Islam, A. Ali, M. Attique, M. Imran, and K.-S. Kwak, "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Future Gener. Comput. Syst.*, vol. 114, pp. 23–43, Jan. 2021.
[5] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Inf. Fusion*, vol. 57, pp. 115–129, May 2020.
[6] M. S. Koti and B. Alamma, "Predictive analytics techniques using big data for healthcare databases," in *Smart Intelligent Computing and Applications*. Cham, Switzerland: Springer, 2019, pp. 679–686.
[7] G. Muhammad, F. Alshehri, F. Karray, A. E. Saddik, M. Alsulaiman, and T. H. Falk, "A comprehensive survey on multimodal medical signals fusion for smart healthcare systems," *Inf. Fusion*, vol. 76, pp. 355–375, Dec. 2021.
[8] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, (Short Papers)*, vol. 2, 2018, pp. 606–611.
[9] V. Kumar, S. Jangirala, and M. Ahmad, "An efficient mutual authentication framework for healthcare system in cloud computing," *J. Med. Syst.*, vol. 42, no. 8, pp. 1–25, Aug. 2018.
[10] A. Ullah, U. Qamar, F. H. Khan, and S. Bashir, "Dimensionality reduction approaches and evolving challenges in high dimensional data," in *Proc. 1st Int. Conf. Internet Things Mach. Learn.*, Oct. 2017, p. 67.
[11] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Inf. Fusion*, vol. 59, pp. 44–58, Jul. 2020.
[12] T. Koç, "A high-dimensional modeling system based on analytical hierarchy process and information criteria," *Math. Problems Eng.*, vol. 2021, pp. 1–9, Aug. 2021.
[13] M. K. Hanif, S. Ayesha, and R. Talib, "Dimension reduction techniques," in *Big Data, IoT, and Machine Learning: Tools and Applications*. Boca Raton, FL, USA: CRC Press, 2021.
[14] W.-Y. Deng, D. Liu, and Y.-Y. Dong, "Feature selection and classification for high-dimensional incomplete multimodal data," *Math. Problems Eng.*, vol. 2018, pp. 1–9, Aug. 2018.
[15] R. Talib, M. Kashif, F. Fatima, and S. Ayesha, "A multi-agent framework for data extraction,Transformation and loading in data warehouse," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 11, pp. 1–4, 2016.
[16] S. C. Ng, "Principal component analysis to reduce dimension on digital image," *Proc. Comput. Sci.*, vol. 111, pp. 113–119, Jan. 2017.
[17] A. Najafi, A. Joudaki, and E. Fatemizadeh, "Nonlinear dimensionality reduction via path-based isometric mapping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1452–1464, Jul. 2016.
[18] M. Muzammal, R. Talat, A. H. Sodhro, and S. Pirbhulal, "A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks," *Inf. Fusion*, vol. 53, pp. 155–164, Jan. 2020.
[19] Z. Wang, L. Qian, C. Han, and L. Shi, "Application of multi-feature fusion and random forests to the automated detection of myocardial infarction," *Cognit. Syst. Res.*, vol. 59, pp. 15–26, Jan. 2020.
[20] S. Ayesha, T. Mustafa, A. Sattar, and M. I. Khan, "Data mining model for higher education system," *J. Sci. Res.*, vol. 43, pp. 24–29, Jan. 2010.
[21] S. Majumder and D. K. Pratihar, "Multi-sensors data fusion through fuzzy clustering and predictive tools," *Expert Syst. Appl.*, vol. 107, pp. 165–172, Oct. 2018.
[22] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
[23] Q. Zhang, C. Yang, D. Wang, Z. Li, Z. Wu, X. Zhu, and Y. Chen, "Atrial fibrillation prediction based on the rhythm analysis of body surface potential mapping signals," *J. Med. Imag. Health Inform.*, vol. 8, no. 1, pp. 145–150, 2018.
[24] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
[25] B. Shin, J. Hogan, A. B. Adams, R. J. Lynch, R. E. Patzer, and J. D. Choi, "Multimodal ensemble approach to incorporate various types of clinical notes for predicting readmission," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, May 2019, pp. 1–4.
[26] V. Vijayasarveswari, A. M. Andrew, M. Jusoh, T. Sabapathy, R. A. A. Raof, M. N. M. Yasin, R. B. Ahmad, S. Khatun, and H. A. Rahim, "Multi-stage feature selection (MSFS) algorithm for UWB-based early breast cancer size prediction," *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0229367.
[27] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Data fusion: Resolving conflicts from multiple sources," 2015, *arXiv:1503.00310*.
[28] M. Baeza, P. Reyes-León, J.-C. Rojo-Mendez, J. Rodríguez-Flores, and E.-G. Perez-Perez, "Data fusion, a multidisciplinary technique," *Int. J. Combinat. Optim. Problems Informat.*, vol. 10, no. 3, pp. 21–32, 2019.
[29] S. Kalamkar and A. G. Mary, "Clinical data fusion and machine learning techniques for smart healthcare," in *Proc. Int. Conf. Ind. 4.0 Technol. (I4Tech)*, Feb. 2020, pp. 211–216.
[30] A. Smirnov and T. Levashova, "Knowledge fusion patterns: A survey," *Inf. Fusion*, vol. 52, pp. 31–40, Dec. 2019.
[31] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Inf. Fusion*, vol. 46, pp. 147–170, Mar. 2019.
[32] Y.-D. Zhang, Z. Dong, S. H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, and F. J. Martinez, "Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation," *Inf. Fusion*, vol. 64, pp. 149–187, Dec. 2020.
[33] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, "Latest trends on heart disease prediction using machine learning and image fusion," *Mater. Today, Proc.*, vol. 37, pp. 3213–3218, Jan. 2021.
[34] S. Liang, X. Deng, and W. Jiang, "Optimal data fusion based on information quality function," *Int. J. Speech Technol.*, vol. 49, no. 11, pp. 3938–3946, Nov. 2019.
[35] A. Abrol, Z. Fu, Y. Du, and V. D. Calhoun, "Multimodal data fusion of deep learning and dynamic functional connectivity features to predict Alzheimer's disease progression," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 4409–4413.
[36] A. Alelaiwi, "Multimodal patient satisfaction recognition for smart healthcare," *IEEE Access*, vol. 7, pp. 174219–174226, 2019.
[37] F. Miao, Z. Liu, J. Liu, B. Wen, Q. He, and Y. Li, "Multi-sensor fusion approach for cuff-less blood pressure measurement," *IEEE J. Biomed., Health Inform.*, vol. 24, no. 1, pp. 79–91, Jan. 2020.
[38] V. Nathan and R. Jafari, "Particle filtering and sensor fusion for robust heart rate monitoring using wearable sensors," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 6, pp. 1834–1846, Nov. 2018.
[39] K. Lin, Y. Li, J. Sun, D. Zhou, and Q. Zhang, "Multi-sensor fusion for body sensor network in medical human–robot interaction scenario," *Inf. Fusion*, vol. 57, pp. 15–26, May 2020.

[40] D. Zhang, Q. Zhu, and D. Zhang, "Multi-modal dimensionality reduction using effective distance," *Neurocomputing*, vol. 259, pp. 130–139, Oct. 2017.
[41] M. U. Khan, S. Aziz, K. Iqtidar, G. F. Zaher, S. Alghamdi, and M. Gull, "A two-stage classification model integrating feature fusion for coronary artery disease detection and classification," *Multimedia Tools Appl.*, pp. 1–30, Apr. 2021.
[42] P. Li, Y. Hu, and Z.-P. Liu, "Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102474.
[43] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–9, Dec. 2020.
[44] Z. Yang, X. Zhuang, C. Bird, K. Sreenivasan, V. Mishra, S. Banks, and D. Cordes, "Performing sparse regularization and dimension reduction simultaneously in multimodal data fusion," *Frontiers Neurosci.*, vol. 13, p. 642, Jul. 2019.
[45] Z. Ahmad, A. Tabassum, L. Guan, and N. M. Khan, "ECG heartbeat classification using multimodal fusion," *IEEE Access*, vol. 9, pp. 100615–100626, 2021.
[46] J. Shi, C. Chen, H. Liu, Y. Wang, M. Shu, and Q. Zhu, "Automated atrial fibrillation detection based on feature fusion using discriminant canonical correlation analysis," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–10, Apr. 2021.
[47] F. Ali, S. El-Sappagh, S. M. R. Islam, D. Kwak, A. Ali, M. Imran, and K.-S. Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Inf. Fusion*, vol. 63, pp. 208–222, Nov. 2020.
[48] S. E. Viswanath, P. Tiwari, G. Lee, and A. Madabhushi, "Dimensionality reduction-based fusion approaches for imaging and non-imaging biomedical data: Concepts, workflow, and use-cases," *BMC Med. Imag.*, vol. 17, no. 1, pp. 1–17, Dec. 2017.
[49] M. M. Hassan, S. Huda, J. Yearwood, H. F. Jelinek, and A. Almogren, "Multistage fusion approaches based on a generative model and multivariate exponentially weighted moving average for diagnosis of cardiovascular autonomic nerve dysfunction," *Inf. Fusion*, vol. 41, pp. 105–118, May 2018.
[50] Q. Cai, H. Wang, Z. Li, and X. Liu, "A survey on multimodal data-driven smart healthcare systems: Approaches and applications," *IEEE Access*, vol. 7, pp. 133583–133599, 2019.
[51] O. Attallah, D. A. Ragab, and M. Sharkas, "MULTI-DEEP: A novel CAD system for coronavirus (COVID-19) diagnosis from CT images using multiple convolution neural networks," *PeerJ*, vol. 8, Sep. 2020, Art. no. e10086.
[52] D. A. Zebari, A. M. Abdulazeez, D. Q. Zeebaree, and M. S. Salih, "A fusion scheme of texture features for COVID-19 detection of CT scan images," in *Proc. Int. Conf. Adv. Sci. Eng. (ICOASE)*, Dec. 2020, pp. 1–6.
[53] H. O. Ilhan, G. Serbes, and N. Aydin, "Decision and feature level fusion of deep features extracted from public COVID-19 data-sets," 2020, *arXiv:2011.08528*.
[54] M. Xu, L. Ouyang, L. Han, K. Sun, T. Yu, Q. Li, H. Tian, L. Safarnejad, H. Zhang, Y. Gao, F. S. Bao, Y. Chen, P. Robinson, Y. Ge, B. Zhu, J. Liu, and S. Chen, "Accurately differentiating between patients with COVID-19, patients with other viral infections, and healthy individuals: Multimodal late fusion learning approach," *J. Med. Internet Res.*, vol. 23, no. 1, Jan. 2021, Art. no. e25535.
[55] A. Jaleel, T. Mahmood, M. A. Hassan, G. Bano, and S. K. Khurshid, "Towards medical data interoperability through collaboration of healthcare devices," *IEEE Access*, vol. 8, pp. 132302–132319, 2020.
[56] A. Tariq, L. A. Celi, J. M. Newsome, S. Purkayastha, N. K. Bhatia, H. Trivedi, J. W. Gichoya, and I. Banerjee, "Patient-specific COVID-19 resource utilization prediction using fusion AI model," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–9, Dec. 2021.
[57] J. Qi, P. Yang, L. Newcombe, X. Peng, Y. Yang, and Z. Zhao, "An overview of data fusion techniques for Internet of Things enabled physical activity recognition and measure," *Inf. Fusion*, vol. 55, pp. 269–280, Mar. 2020.
[58] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Inf. Fusion*, vol. 50, pp. 158–167, Oct. 2019.
[59] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Inf. Fusion*, vol. 53, pp. 209–221, Jan. 2020.
[60] R. C. King, E. Villeneuve, R. J. White, R. S. Sherratt, W. Holderbaum, and W. S. Harwin, "Application of data fusion techniques and technologies for wearable health monitoring," *Med. Eng. Phys.*, vol. 42, pp. 1–12, Apr. 2017.
[61] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," 2021, *arXiv:2102.01998*.
[62] M. M. Alyannezhadi, A. A. Pouyan, and V. Abolghasemi, "An efficient algorithm for multisensory data fusion under uncertainty condition," *J. Electr. Syst. Inf. Technol.*, vol. 4, no. 1, pp. 269–278, May 2017.
[63] Q. He, X. Li, D. W. N. Kim, X. Jia, X. Gu, X. Zhen, and L. Zhou, "Feasibility study of a multi-criteria decision-making based hierarchical model for multi-modality feature and multi-classifier fusion: Applications in medical prognosis prediction," *Inf. Fusion*, vol. 55, pp. 207–219, Mar. 2020.
[64] Kaggle. (2020). *COVID-19 X-Ray Images*. [Online]. Available: https://www.kaggle.com/bachrr/covid-chest-xray
[65] K. Dramane, G. B. Tra, K. K. Prosper, and I. C. Yamoussoukro, "New hybrid method for efficient imputation of discrete missing attributes," *Int. J. Innov. Appl. Stud.*, vol. 31, no. 4, pp. 763–775, 2021.
[66] G. Texier, R. S. Allodji, L. Diop, J.-B. Meynard, L. Pellegrin, and H. Chaudet, "Using decision fusion methods to improve outbreak detection in disease surveillance," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–11, Dec. 2019.
[67] J. Kittler and F. Roli, *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21-23, 2000 Proceedings*. Berlin, Germany: Springer, 2003.
[68] I. H. Sarker, A. S. M. Kayes, and P. Watters, "Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage," *J. Big Data*, vol. 6, no. 1, pp. 1–28, Dec. 2019.
[69] F. Fatima, R. Talib, M. K. Hanif, and M. Awais, "A paradigm-shifting from domain-driven data mining frameworks to process-based domain-driven data mining-actionable knowledge discovery framework," *IEEE Access*, vol. 8, pp. 210763–210774, 2020.
[70] P. Shi, G. Li, Y. Yuan, and L. Kuang, "Data fusion using improved support degree function in aquaculture wireless sensor networks," *Sensors*, vol. 18, no. 11, p. 3851, Nov. 2018.
[71] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, "Text mining: Techniques, applications and issues," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 11, pp. 414–418, 2016.
[72] P. Potapov and A. Lubk, "Optimal principal component analysis of STEM XEDS spectrum images," *Adv. Structural Chem. Imag.*, vol. 5, no. 1, pp. 1–21, Dec. 2019.