Supporting Multidimensional Risk Analysis and Mining over Big Digital Twins: An Effective and Efficient Framework

Alfredo Cuzzocrea [1,2,*], Ismail Benlaredj [1]

[1] *iDEA Lab, University of Calabria, Rende, Italy*
[2] *Dept. of Computer Science, University of Paris City, Paris, France*

Abstract

The emergence of *Digital Twin technology*, coupled with *Big Data paradigms*, has enabled advanced analytics for *real-time monitoring*, *simulation*, and *decision-making* across multiple domains. In this paper, we address the emerging challenge of *multidimensional risk analysis and prediction* in Digital Twin environments, by proposing an innovative *Multidimensional Big Data Analytics framework* that integrates *Digital Twin modeling*, *Cloud-based infrastructures*, and *OLAP-driven multidimensional analysis* to assess and predict risks effectively. Our approach is designed to support complex risk scenarios, such as those in healthcare, by leveraging distributed data processing, real-time analytics, and predictive modeling. Through extensive experimental evaluations over real-life healthcare datasets, we demonstrate the *efficiency* and *scalability* of our framework in performing multidimensional risk analytics and prediction.

Keywords[1]

Digital Twins, Big Data, Multidimensional Risk Analysis, Multidimensional Risk Prediction

1. Introduction

Digital twins (e.g., [1-3]) have emerged as a transformative technology in the era of *big data* (e.g., [4-6]). By integrating real-time data from sensors and historical information, digital twins enable the simulation, monitoring, and optimization of physical assets across various domains. The convergence of digital twin technology with big data applications has opened new avenues for research, focusing on several key aspects. Firstly, the *data integration and management* aspect is crucial. Digital twins rely on vast amounts of data from diverse sources, necessitating advanced data management techniques to ensure data quality, consistency, and accessibility. Researchers are exploring innovative methods to handle the volume, velocity, and variety of big data, including the use of cloud computing and edge computing to enhance data processing capabilities. Secondly, the *modeling and simulation* aspect involves creating accurate and dynamic models of physical assets. This requires sophisticated algorithms and *Machine Learning* (ML) techniques to predict and replicate the behavior of these assets under various conditions (e.g., [7,8]). Research in this area aims to improve the fidelity of simulations and the ability to perform real-time updates based on incoming data. Thirdly, the *real-time analytics and decision-making* aspect is pivotal. Digital twins enable real-time monitoring and predictive maintenance by analyzing data streams to detect anomalies and predict failures before they occur (e.g., [9]). This aspect of research focuses on developing robust analytics frameworks and decision-support systems that can provide actionable insights in real-time. Lastly, the *interoperability and standardization* aspect addresses the need for standardized protocols and frameworks to ensure seamless integration and communication between digital twins and other systems (e.g., [10,11]). Researchers are working on establishing common standards and best practices to facilitate the widespread adoption and scalability of digital twin technology.

[1] *SEDB 2025: 33rd Symposium on Advanced Database Systems, June 16-19, 2025, Ischia, Italy*
* This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at University of Paris City, Paris, France
* Corresponding author.
alfredo.cuzzocrea@unical.it (A. Cuzzocrea); ismail.benlaredj@unical.it (I. Benlaredj)
: 0000-0002-7104-6415 (A. Cuzzocrea); 0009-0003-1138-8039 (I. Benlaredj)

In summary, the research on digital twins for big data applications is multifaceted, encompassing data management, modeling and simulation, real-time analytics, and interoperability. These efforts are driving the evolution of digital twins from conceptual models to practical tools that enhance efficiency, reliability, and decision-making across various industries.

In this paper, we focus on a specific, innovative research challenge: *risk analysis over digital twins* (e.g., [12,13]). The final goal of this new idea is to design and develop data analytics methodologies for characterizing, analyzing and, moreover, making predictions about the *risk of organizations* (e.g., healthcare systems) by inspecting the data and knowledge kept in digital twins. Without any loss of generality, big data technologies have largely been used for risk management activities, in various organizations (e.g., [14,15]). Given the strict interaction with *data analytics methodologies* and ML, several methodologies have been applied for analyzing organizational risk using digital twin data, particularly in sectors like healthcare. For instance, techniques such as *regression analysis*, *decision trees*, and *neural networks* are used to predict potential risks by analyzing historical and real-time data from digital twins [14]. For instance, in healthcare, predictive analytics can forecast *patient deterioration* or *equipment failure*. Similarly, *time series analysis* analyzes data points collected or recorded at specific time intervals to identify trends, seasonal patterns, and potential anomalies that could indicate risks [15]. Tools like *dashboards* and *heat maps* help in visualizing complex data from digital twins, making it easier to identify *patterns* and *outliers* that could signify risks [16]. *Descriptive statistics* such as mean, median, and standard deviation are used as well to summarize and describe the main features of a dataset, providing insights into the normal operating conditions and deviations that might indicate risks [17]. Indeed, leveraging digital twin data through various data analytics methodologies allows organizations to proactively identify, assess, and mitigate risks. These methodologies enhance decision-making, improve operational efficiency, and ultimately contribute to better risk management practices.

In line with this relevant research trend, this paper proposes the application of the innovative *multidimensional big data analytics paradigm* [18-20] to the specific case of risk analysis, via proposing an innovative model, and the related framework, for *supporting multidimensional risk analysis and prediction over digital twins*.

The remaining of this paper is structured as follows. Section 2 offers a comprehensive overview of state-of-the-art proposals related to the context of multidimensional risk analysis and prediction over digital twins. Section 3 summarizes our proposed multidimensional big data analytics paradigm as applied to the specific research context. Section 4 describes the framework that implements our methodology for multidimensional risk analysis and prediction. In Section 5, we introduce and describe the real-life datasets used in our experimental assessment. Section 6 presents the first part of our experimental campaign, which consists of multidimensional risk analytics. Similarly, Section 7 presents the second part, which consists of the multidimensional risk prediction. Finally, Section 8 provides conclusions and possible future work.

2. Related Work

In this Section, we provide a comprehensive overview of state-of-the-art proposals in the context of multidimensional risk analysis and prediction over digital twins.

[21] contributes significantly to the intersection of *computational intelligence*, *data privacy* and *Digital Twins* (DT) technology for *Cyber-Physical Systems* (CPS) in smart cities. The approach centers on using *Differential Privacy Frequent Subgraph-Big Multigraph* (DPFS-BM) for preserving data privacy in DT, which maps the complex physical spaces of CPS into virtual environments. Authors propose a novel *Differential Privacy-AlexNet* (DP-AlexNet) model for analyzing and predicting *Big Graphic Data* (BGD) in DT, solving the gradient dispersion problem typically encountered in deep networks. Through comparative analysis with various neural networks (e.g., CNN, RNN, etc.), the model demonstrates superior performance in terms of error rates, training, and testing time.

In the field of precision medicine for *Cardio-Vascular Disease* (CVD), the concept of digital twins has gained attention for its potential to improve *patient-specific diagnosis* and treatment. By creating a virtual representation of a patient that integrates real-time updates from various *clinical*, *imaging*,

molecular and so forth, the DT can optimize treatment decisions and predict disease progression. In [22], authors conduct a mapping review of DT research in CVD. The review also highlighted the growing commercial interest in DT technology, with several companies that focus on *CVD-related digital twin models*, including simulation models that act as precursors to real-time cyber-physical systems. Despite the promising interdisciplinary and global nature of this research, authors identified several challenges, such as *ethical concerns* and *clinical barriers* to the adoption of *AI-driven decision* tools in healthcare.

[23] presents a comprehensive review of the application of ML techniques in healthcare, particularly focusing on health prediction systems using *Electronic Health Record* (EHR) datasets. Authors emphasize the importance of ML in automating healthcare processes and improving disease prediction accuracy. By leveraging ML methods, this review highlights the potential for intelligent systems to learn from past experiences, by enabling more accurate and proactive healthcare solutions. Authors argue that ML can extract valuable insights from EHRs, ultimately contributing to better health risk prediction. This work serves as a foundational resource for researchers looking to explore the intersection of ML and healthcare prediction.

In [24] authors explore the application of a *Fuzzy Control System* (FCS) in healthcare for diagnosing human health risks related to blood pressure, and blood glucose across different age groups. The FCS is recognized for its ability to handle uncertainty in clinical decision-making and treatment, which makes it suitable for addressing the variability in medical diagnoses. This research predicts health risks by employing FCS and compares the results with previous studies. The system supports reliable decision-making, by enabling immediate treatment recommendations and lifestyle adjustments to mitigate future health risks. Authors also suggest that this FCS approach can be extended to predict additional health risks, including mental and physical conditions, providing a versatile tool for healthcare diagnostics.

[25] explores the challenges and opportunities of leveraging *Electronic Health Records* for *data-driven healthcare*. Given the inherent difficulties in working with EHR data, such as *temporality*, *sparsity*, and *noise*, authors propose a *Deep Learning-based approach* for phenotyping from patient EHRs. Their method represents each patient's EHR as a temporal matrix, where time and medical events form the two dimensions. A four-layer *convolutional neural network* model is then applied. To account for temporal dynamics, the study investigates three fusion techniques: *early*, *late*, and *slow fusion*. Moreover, the proposed model is validated over a real-life EHR dataset, particularly in the context of predictive modeling for *chronic diseases*. This evaluation demonstrates its effectiveness in phenotype extraction and disease prediction.

In [26], authors explore the impact of *Digital Twin technology* on healthcare, by emphasizing its transformative potential in *diagnostics*, *drug delivery*, and treatment optimization as part of *digital transformation* adopted by *Industry 4.0*. Digital Twin technology has broad applications in healthcare, from hospital operations and public health processes and so forth. Motivated by the COVID-19 pandemic, authors focus on applying DT technology for *virus containment* in workplaces through social distancing measures. Their contributions consist of introducing a generalized DT architecture to identify key functional components of DT systems and then presenting "*CanTwin*" an innovative case study that utilizes DT technology in a *canteen service* for monitoring social distancing. This study provides valuable insights into how DT technology can be applied for real-time monitoring and optimization in healthcare.

[27] highlights the promising role of digital twin technology in healthcare, which focuses on its applications in *monitoring*, *diagnosis*, *treatment development* and *drug discovery*. Digital twins, acting as virtual counterparts of real human patients, leverage various data sources such as blood glucose levels, heart imaging, cardiac electrophysiology, and multi-omics data. The article explores the challenges of standardizing, integrating, and interpreting these diverse datasets and highlights how different methods help to overcome these issues in order to create effective digital twins. Despite the significant progress in DT technology, authors note that further advancements in *non-invasive data collection*, as a result, modeling techniques and computational power are needed to achieve fully comprehensive patient digital twins.

[28] presents a literature review on the use of digital twin technology in *safety analysis*, *risk assessment* and *emergency management*. While digital twins are emerging as powerful tools for

monitoring complex systems, autonomous control, and real-time assistance during emergencies, authors emphasize that key issues such as safety, cybersecurity, and reliability remain unresolved. These concerns are crucial for evaluating the risks and benefits associated with DT implementation. This research has two main objectives, (*i*) it reviews the latest advancements in DT technology, by offering a catalog of expected functions and enabling technologies across various application domains; (*ii*) it highlights the limitations and challenges that still need to be addressed for effective digital twin integration in safety and risk management scenarios. This study underscores the importance of further research to enhance the safety and reliability of DT, particularly in *critical* and *high-risk environments*.

[29] provides a review of digital twin technology, focusing on its role in *smart manufacturing* and *Industry 4.0*. DT facilitates unified integration between the *cyber* and *physical* spaces, and their significance is increasingly recognized in both *academic* and *industrial* literature. Despite the concept being introduced nearly 15 years ago and its successful application in various industries (e.g., product design, production, and health management). This paper addresses those concerns by reviewing the key components, current developments, and major applications of DT in industry. Additionally, it discusses ongoing challenges and proposes future research directions to advance DT implementation in industrial environments.

Finally, [30] presents a novel approach that combines *reduced-order models* with ML to create physics-informed digital twins for *nuclear reactor core simulations*. The digital twin predicts complex outputs like neutron flux and power distribution by solving forward and inverse problems. *Proper orthogonal decomposition* is used offline to build accurate computational models, while ML techniques, such as *k-Nearest-Neighbors* (KNN) and *decision trees*, handle *input-parameter-dependent* coefficients.

From the analysis of the active literature, the strident interest for our investigated topic clearly emerges.

3. Multidimensional Big Data Analytics Applied to the Multidimensional Risk Analysis and Prediction Problem

In this Section, we propose a brief summary about the multidimensional big data analytics paradigm we proposed recently [18-20], and how it is applied to the multidimensional risk analysis and prediction problem.

Basically, the proposed methodology aims at applying the well-known multidimensional and OLAP analysis paradigm [31] to the *core* of emerging big data analytics procedures (e.g., [32]), even within the applicative context of state-of-the-art big data processing platforms, such as *Apache Hadoop* and *Apache Spark*. The methodology relies on the following main steps:

- designing and implementing a proper *big multidimensional analysis model* over the target heterogeneous data sources, driven by specific analysis goals;
- designing and implementing a proper (*distributed*) *MapReduce-based big multidimensional data storage model*, based on a reference big data platform (e.g., Apache Hadoop);
- designing and implementing a proper collection of (*distributed*) *functional-based multidimensional big data analytics procedures*, driven by specific analytics goals;
- designing and implementing a proper collection of (*distributed*) *multidimensional big data visualization tools*, driven by the target knowledge fruition .

It should be noticed here that, as regards research challenges, the most annoying problem is represented by the issue of creating and developing ad-hoc *big data management and processing algorithms*, which support the analytics and prediction phases, by strictly taking into account *performance aspects*, which play a relevant role in emerging big data research (e.g., [33,34]). Indeed, computational overheads are a challenging aspect of these issues, when applied to real-life big data settings, so that appropriate solutions must be taken.

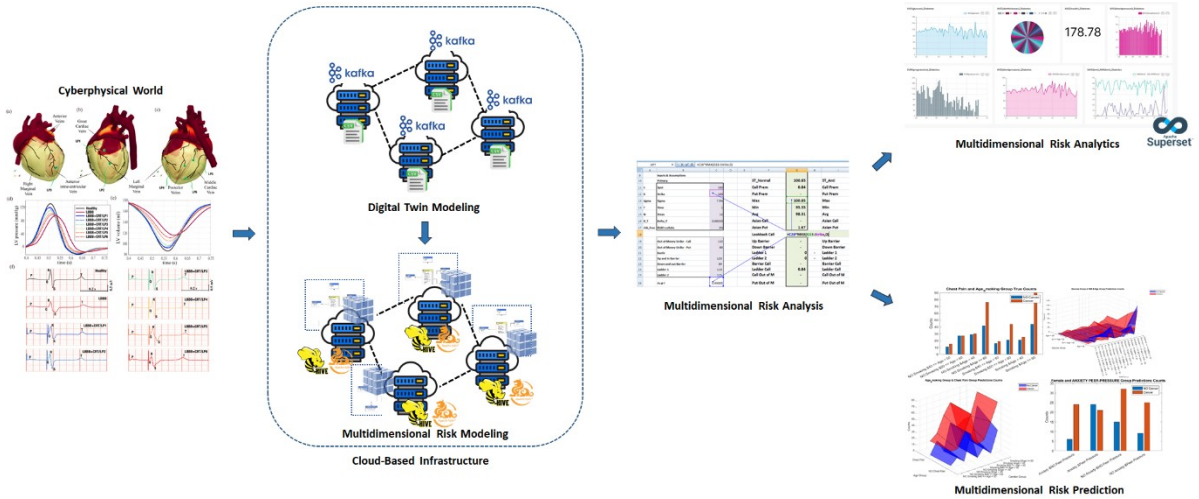


Figure 1: The Reference Framework: Supporting Multidimensional Risk Analysis and Prediction over Digital Twins

When applied to the big data scenarios represented by digital twins, the most significant aspect for the methodology we propose is represented by the issue of *appropriately capturing the streaming and pervasive nature of such kind of data sources*, by integrating a proper functional component. In addition, another important aspect for the case of digital twins is represented by *the need for big data prediction tools*, due to specific operational settings of these environments.

Finally, still in the specific context we investigate, the core multidimensional model is "shaped" as a multidimensional risk model, meaning that the measures to analyze are connected to the evaluation of the risk in the specific setting (e.g., healthcare) and the core multidimensional predictive model is "shaped" as a multidimensional risk prediction model, both tailored to digital twins.

4. An Innovative Framework for Supporting Multidimensional Risk Analysis and Prediction over Digital Twins

In this Section, we introduce our proposed innovative framework designed to support multidimensional risk analysis and prediction over digital twins, particularly in the healthcare domain. Notably, this framework is distinguished by its integration of *digital twin modeling*, *Cloud-based infrastructure*, and advanced analytics, which enable risk assessment and prediction. Figure 1 shows a comprehensive overview of this reference framework.

As shown in Figure 1, the framework is composed of several key layers, each playing a crucial role in the overall functioning of the framework. These layers are described below and will be discussed in detail in the following.

- *Cyberphysical World*: this layer focuses on gathering data from various sensors, such as *heart sensors*, which provide critical information about heart disease analysis;
- *Cloud-Based Infrastructure*: this infrastructure consists of two core components: first, *Digital Twin Modeling* component, which involves interconnected Cloud nodes with *Apache Kafka instances* that handle streaming data produced by digital twins; second, *Multidimensional Risk Modeling* component is responsible for creating data cubes in each Cloud node in order to be efficiently processed and analyzed;
- *Multidimensional Risk Analysis*: this layer performs Cloud-based OLAP-shaped multidimensional risk analysis, enabling the exploration of heart disease risks from multiple dimensions;

- *Multidimensional Risk Analytics*: this layer focuses on supporting multidimensional risk analytics derived from the analysis, making it easier to interpret and understand the multidimensional risk data;
- *Multidimensional Risk Prediction*: this final layer is responsible for providing predictive insights based on multidimensional risk analysis.

In the following, we examine these layers and components with greater detail, by highlighting their specific roles within the proposed framework. While our proposed framework is general enough to capture any innovative big data setting scenario, here we specialize our description for the very relevant case represented by the *digital twin healthcare scenario* (e.g., [35]).

1. *Cyberphysical World*: this layer is the initial layer of our framework, where real-time healthcare data from the physical environment are captured. In the context of *cardiovascular healthcare*, it includes patient health metrics such as *heart rate*, *blood pressure*, *Electrocardiogram* (ECG) readings, and other cardiovascular signals. The aim of this layer is to ensure that an accurate dataset is available for analysis. The data gathered here serves as input to the following layers in the framework. This layer is crucial because the accuracy and quality of data directly impact the effectiveness of the analytical models and predictions of heart diseases;
2. *Cloud-Based Infrastructure*: this layer is crucial due to its importance in providing an environment for processing and managing large amounts of data collected from the cyberphysical world layer. The infrastructure is composed of two primary components, (*i*) *Digital Twin Modeling* component; (*ii*) *Multidimensional Risk Modeling* component, each playing a pivotal role in the data analysis process.
    a. *Digital Twin Modeling*: this component is composed of several interconnected Cloud nodes, each equipped with Apache Kafka instances to manage the data streams of the collected data from heart sensors. Later the collected data is transmitted in CSV format and fed into these Cloud nodes through Apache Kafka;
    b. *Multidimensional Risk Modeling*: the second critical part of the Cloud-based infrastructure is the Multidimensional Risk Modeling component, which transforms data into *multidimensional models* (OLAP data cubes – e.g., [31,36]). Each data cube is created and maintained in a distributed-Cloud-mode (e.g., [37,38]). With this distributed architecture, the efficiency of data processing is enhanced by enabling parallel operations across nodes. The multidimensional nature of these cubes allows for OLAP-based querying (e.g., [39]), which is essential for extracting and analyzing the relationships and trends of data. The OLAP cubes generated by the Cloud nodes are then materialized within *Apache Hive*. This step allows us to integrate the data into *Apache Kylin*, which is in charge of supporting efficient OLAP querying and multidimensional analysis. Then, we use *Apache Superset* to visualize the multidimensional data in an intuitive format by connecting to Apache Kylin and retrieving the multidimensional data. This setup provides a robust platform for visualizing and exploring healthcare analytics.
3. *Multidimensional Risk Analysis*: this layer focuses on processing aggregate data from the Cloud infrastructure and deriving meaningful insights through multidimensional querying and exploration, thus fully realizing the multidimensional risk analytics phase. This phase involves organizing data into dimensions and measures, such as *cardiovascular metrics* (e.g., *heart rate*, *blood pressure*, and so forth) and *patient demographics* categorized by *time*, *age group*, and *region* for efficient analysis. The risk analysis algorithms then apply transformations, aggregations and statistical analyses to identify trends, patterns and correlations within the data. Concurrently, a continuous flow of data between various fragments of this risk analysis layer enables multidimensional querying. Users can drill-down into specific data or aggregate data across various dimensions, providing deeper insights into *cardiovascular risk*;
4. *Multidimensional Risk Analytics*: it focuses on translating complex analytical results into *easy-to-understand* visual representations for healthcare providers to interpret the results and make informed decisions. By employing a variety of visualization tools such as *bar plots*, *pie plots*, *3D*

plots, and so forth, the framework allows users to explore and analyze risk effectively. These visual tools help users gain a deeper understanding of the relationships between different risk factors and patterns emerging from multidimensional data analysis;

5. *Multidimensional Risk Prediction*: the final component of the framework focuses on multidimensional risk prediction, which translates the analyzed data into actionable predictions. This layer is responsible for generating predictions based on the processed healthcare data, using ML models and statistical tools. After generating predictions, the next phase consists of presenting the results in an intuitive manner through the usage of plots and graphs. The 3D plots and bar plots show trends and relationships between risk factors, while comparison plots illustrate the likelihood of specific outcomes (e.g., *chest pain prevalence*, *glucose level predictions*, etc.).

Finally, the proposed framework operates in a *continuous cycle*, starting with real-time data collection from the *cyberphysical* world, which is then mirrored virtually via digital twin modeling. These data flow through Cloud-based infrastructure for risk modeling and culminate in visualized predictions and risk assessments. The seamless integration of these technological components allows healthcare providers to monitor patient conditions in real-time, predict future risks, and take preventive actions accordingly. This comprehensive solution effectively addresses the challenges faced by modern healthcare systems.

5. Experimental Assessment and Analysis: Dataset Description

In this Section, we introduce the following three *healthcare-related* datasets: *HeartStudy Dataset* [40], *DiabetesHealthcare Dataset* [41] and *LungCancer Dataset* [42]. These datasets have been used in our experimental campaign. We explore the specific attributes that each dataset contains, and also examine the structure of these datasets in detail discussing the key variables that contribute to understanding heart disease, diabetes and lung cancer. As mentioned in Section 4, data contained in these datasets have been used to feed the Apache Kafka component, like they were produced by the digital twin layer of the framework. In turn, Apache Kafka aliments the Cloud storage of the framework, and these Cloud-enabled datasets are finally used to populate the OLAP data cubes that are built according to the multidimensional big data analytics paradigm [18-20].

HeartStudy [40] is a dataset about heart disease that includes 4,240 records and 16 columns, with 15 attributes related to patient information. This dataset aims to predict whether a patient has a *10-year* risk of developing *Coronary Heart Disease* (CHD). The dataset target attribute named *TenYearCHD* is a binary indicator representing the 10-year risk of CHD. It takes the value 1 if the patient is at risk of developing CHD in the next 10 years, or 0 otherwise.

DiabetesHealthcare [41] is a dataset sourced from the *National Institute of Diabetes and Digestive and Kidney Diseases* (NIDDK) project that focuses on creating knowledge and treatments for chronic, costly, and consequential diseases. The objective is to predict whether or not a patient has *diabetes* based on various diagnostic measurements. The dataset includes 8 medical variables and one target variable named *Outcome* indicating whether the patient has diabetes or not (i.e., 1 or 0).

LungCancer [42] is a dataset that focuses on evaluating the effectiveness of a cancer prediction system by providing information to assess lung cancer risk at a low cost. The data is collected from an *Online Lung Cancer Prediction System. Gender* attribute has two values *M* for *Male* and *F* for *Female.* All other attributes have two values 1 or 2 (e.g., *Smoking* attribute the value 1 indicates that the person is not smoking, and 2 indicates the person is smoking).

6. Experimental Assessment and Analysis: Multidimensional Risk Analytics

In this Section, we are going to experiment with three different datasets related to healthcare, with particular attention to the multidimensional risk analytics supported by our proposed framework. Accordingly, we create dashboards for each dataset using Apache Superset in order to present key analytics and visualizations that help uncover critical patterns, trends, and actionable insights,

Figure 2: *HeartStudy* Dashboard

contributing to data-driven decision-making in healthcare.

As described in Section 4, the multidimensional risk analytics is generated by our multidimensional OLAP model embedded in the framework, which adheres to the underlying multidimensional big data analytics paradigm [18-20].

First, we consider the *HeartStudy* dataset. The constructed *HeartStudy dashboard* showcases a series of analytical plots related to the target dataset. These visualizations provide insights into various key metrics and patterns, enabling a deeper understanding of the dataset attributes and trends. Figure 2 shows the *HeartStudy dashboard*, which provides an overview of important smoking and health patterns. The dashboard focuses the attention on the relationship between smoking cigarettes and CHD risk. It demonstrates 37.9K cigarettes per day in the overall dataset, and the most represented age in the dataset was from 40 to 50. Males smoke more than females. The cigarette smoking rate for people aged 40 to late 40s is the highest. Both males and females have roughly the same average *glucose* level. At the age of 32, smoking rates reached their peak and remained consistent thereafter.

Second, we consider the *DiabetesHealthcare* dataset. Figure 3 shows the *DiabetesHealthcare dashboard*, which is underlined by a few of the following health metric indicators. The AVG(*glucose*) maintains its stability in the range between 80 to 160. The histogram of SUM(*pregnancies*) shows that women aged 40 to 45 have the highest number of pregnancies. The area plot of AVG(*bloodpressure*) shows stability in the range of 50 to 90. The pie plot of AVG(*skinthickness*) shows similar averages across all ages. The AVG(*insulin*) is 178.78. Finally, two line plots represent how the MAX(*BMI*) will be in the range of 40 to 65, and MIN(*BMI*) is in the range of 0 to 30.

Finally, the *LungCancer* dataset is considered, being the corresponding *LungCancer dashboard* shown in Figure 4. Here are some of the insights derived from this dashboard. Two pie plots reveal that females have slightly higher average anxiety levels, while the COUNT(*lung_cancer*) is relatively the same for both genders. An histogram of COUNT(*smoking*) shows that males smoke more than females, with the most smokers in the late 50s to early 60s age group. Another histogram of AVG(*allergy*) indicates stability, with males slightly more affected. A line plot of AVG(*anxiety*) shows that both genders have similar anxiety levels, and the total COUNT(*smoking*) is 1.01K.

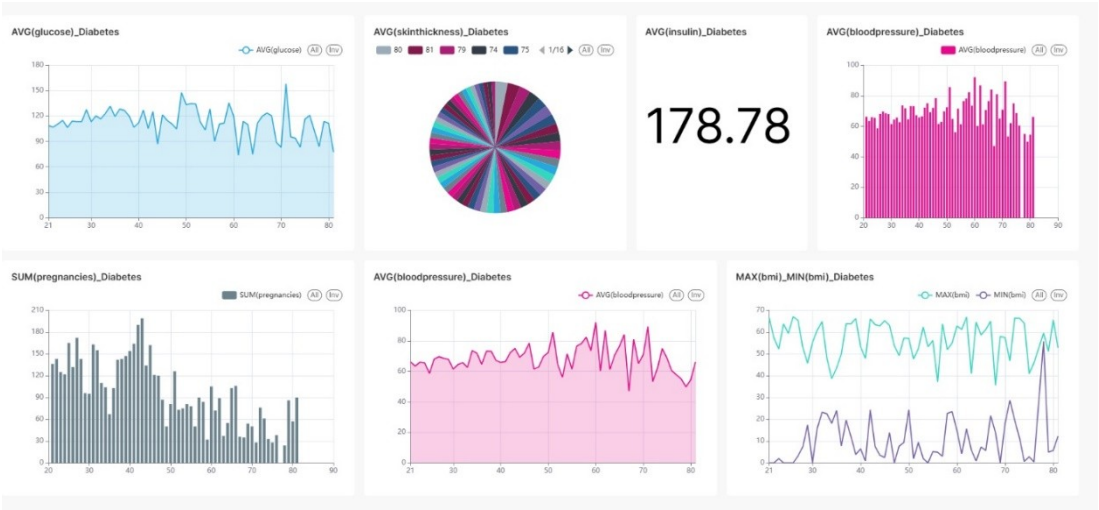


Figure 3: *DiabetesHealthcare* Dashboard

7. Experimental Assessment and Analysis: Multidimensional Risk Prediction

In this Section, we focus on the second aspect of our experimental assessment and analysis, i.e. the multidimensional risk prediction. Our model includes all steps from data processing, analysis, and generating predictions to the final results, and advanced techniques including data aggregation, multidimensional data analysis, and reporting with suitable plots. The very relevant innovation of our risk prediction approach is, again, the *multidimensionality of the analysis*, meaning that our approach completely captures the multidimensional nature of the surrounding analysis problem, in order to finally enhance the *expressive power* of the overall (analytics) process.

In our experimental campaign, we have developed *Credit Scoring* [43], which is normally utilized by financial institutions as a statistical tool to decide the *creditworthiness* of either an individual or an organization. However, our *customized* model is used for *healthcare-related* multidimensional prediction from digital twin data. Indeed, instead of financial dimensions, our custom model evaluates variables like *medical history*, *clinical indicators*, *chronic conditions*, *lifestyle choices*, and various health measurements to predict the likelihood of future health issues or events. The second prediction model is the *Deep Learning Network* [44] used for estimating the *Probability of Default* (PD). Basically, this model trains a *credit risk* for PD prediction using a *Deep Neural Network* [45]. The model focuses on alternative network designs and fits simpler models without *macroeconomic variables*. Again, like in the case of *Credit Scoring*, we customize the model implementation as to support multidimensional prediction from digital twin data, via proper multidimensional healthcare model.

In the following, due to space limitation, we present our experimental results as restricted to the multidimensional risk prediction over the *HeartStudy* dataset only, but we conducted the prediction accuracy analysis over all the three datasets, with similar results.

7.1. Credit Scoring Model

In our *Credit Scoring* model [43], we use a *credit ScoreCard* function to generate predictions based on specific health metrics. This allows us to predict health risks based on critical patient attributes. Subsequently, we use a *Fitmodel* function to fit a *Logistic Regression Model* to the *Weight of Evidence* (WOE) data. *Fitmodel* internally bins the training data, and transforms it into WOE values,

Figure 4: *LungCancer* Dashboard

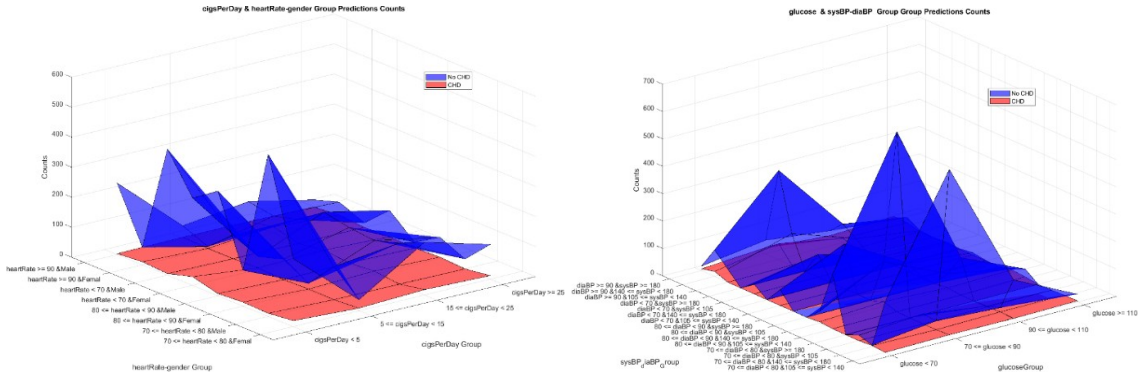
after that, we compute the *probability of default* for the *credit scorecard model* using the *Probdefault* function, and we define the threshold for the *probability of default* as 0.35. Finally, we use custom functions for the visual reports to plot aggregate results, and we use $3D$ plots for more detailed reports.

Here, we present the results of applying our *Credit Scoring* model to the first of the three healthcare datasets previously described (i.e., *HeartStudy* dataset). By focusing on this initial dataset, we provide a closer look at the predictions, combining and aggregating various attributes. This approach uncovered valuable insights and helped identify relationships between key indicators.

Through the application of our model on the *HeartStudy* dataset, we generate predictions for the target variable, which is a 10-year risk of future CHD, we provide detailed analysis and informative reports. These outputs are visually represented in $3D$ plots.

Figure 5 displays the *prediction counts* for CHD. In order to capture the multidimensional modelling, in Figure 5 (a) we select three attributes, including the target variable CHD. Numerical variables, such as *heartRate* and *cigsPerDay*, have been grouped into intervals (mostly four intervals). The first two attributes, *heartRate* and *gender*, have been combined resulting in a combination of intervals, while the third attribute, *cigsPerDay*, has been presented independently. The results have been aggregated based on the three selected attributes, with the prediction counts split into two categories: *CHD* and *No CHD*. Overall, this originates a multidimensional prediction model over digital twin data extracted from the target dataset, according to the three-dimensional range ⊠*cigsPerDay*, *heartRate*, *gender*⊠.

Similarly, Figure 5 (b) shows the prediction counts for CHD over the range ⊠*glucose*, *diaBP*, *sysBP*⊠. For this range, we selected the three attributes, *glucose*, *diaBP* and *sysBP*, including the target variable CHD. All these attributes were grouped into four intervals. The first two attributes, *sysBP* and *diaBP*, have been combined resulting in a combination of bins, while the third attribute, *glucose*, has been used independently. It is easy to notice, here, that this approach can easily scale to higher numbers of dimensions, since multiple dimensions can be easily grouped/projected into different sub-groups of dimensions (e.g., [44]).



(*a*) (*b*)

Figure 5: Prediction Counts for $\langle cigsPerDay, heartRate, gender \rangle$ (*a*) and $\langle glucose, sysBP, diaBP \rangle$ (*b*) (*Credit Scoring* Model)

As shown in Figure 5 (b), the predictions demonstrate that patients with high values of the three attributes ($glucose > 110$, $diaBP > 90$, and $sysBP > 180$) have a serious risk of developing CHD. We observe the risk of having CHD decreases with *glucose* less than 110; however, we notice that high *diaBP* and *sysBP* values still put some patients at risk of developing CHD even with a low level of *glucose*.

Figure 6 shows the prediction counts for CHD over the range ⟨*currentSmoker*, *heartRate*, *prevalentHip*⟩. In this case, we selected three attributes, including the target variable CHD. All these attributes were grouped into four intervals. The first two attributes, (i.e., *currentSmoker* and *prevalentHip*), have been combined while the third attribute, *heartRate*, has been represented independently.

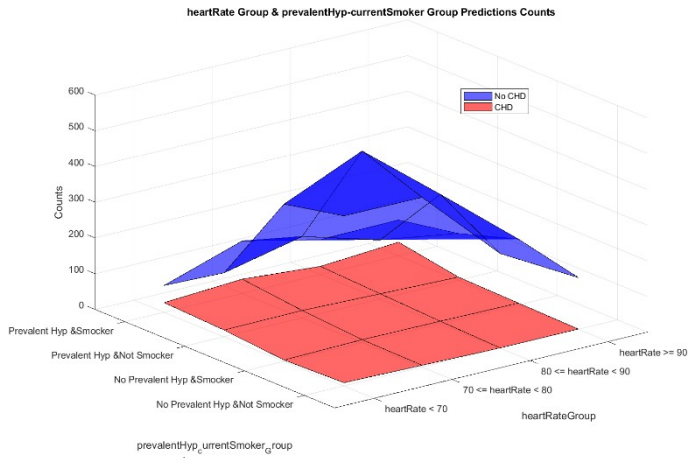


Figure 6: Prediction Counts for $\langle currentSmoker, heartRate, prevalentHyp \rangle$ (*Credit Scoring* Model)

As a result, the predictions show that patients who are smokers and who have prevalent hypertension are at risk of developing CHD, especially patients with irregular heart rates ($heartRate > 90$). On the other hand, predictions show that non-smoker patients without prevalent hypertension have a low risk of CHD.

7.2. Deep Learning Network Model

In our *Deep Learning Network* model [44], we train a *credit risk for the probability of default* prediction using a Deep Neural Network. The model starts by loading and preprocessing the data, all attributes

should be appropriately formatted to fit the input of the model. After preparing the data, we divide it into *training*, *validation*, and *testing* sets. The training data is used to build the model, the validation data is utilized to fine-tune and confirm the model parameters, and the testing data is used to check the overall performance of the models.

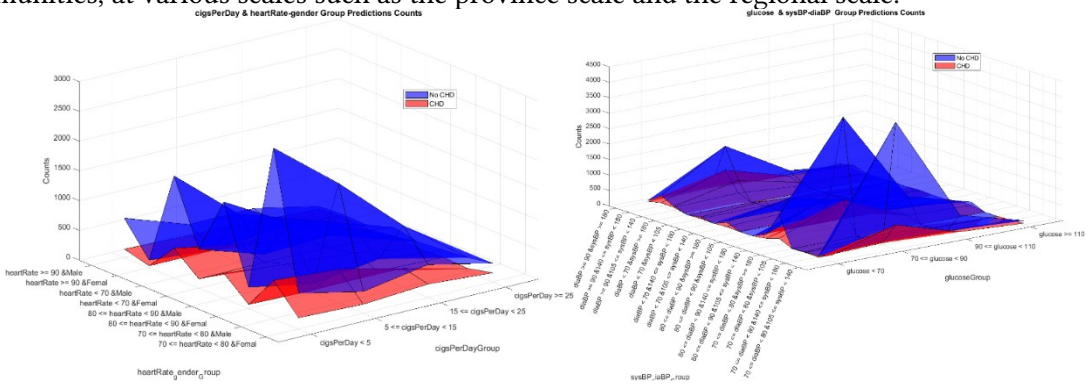
With prepared data, we define the network architecture, including the parameters of layers, functions, and other key components. Afterward, we fine-tune the training parameters, such as *Learning Rate*, *Batch Size* and *Epochs Number*. Once these steps are completed, we proceed to train the network using the prepared data. For efficiency, a pre-trained model, named *residualTrainedNetworkMacro*, is provided, which allows us to *bypass* the training process and directly explore the results with other versions of the dataset. Moreover, the same advanced visualization techniques are employed for more detailed reports.

Through the application of our model on the *HeartStudy* dataset, we generate predictions for the target variable, which is the 10-year risk of future coronary heart disease CHD, thus furnishing detailed analyses and informative reports. These outputs are then visually represented in ad-hoc plots. The prediction model discriminates between the two same categories exploited for the *Credit Scoring* model, i.e. *CHD* and *No CHD* (see Section 7.1).

Figure $7\ (a)$ displays the prediction counts for CHD over the range $\langle cigsPerDay,\ heartRate,\ gender \rangle$. By analyzing the results, the predictions indicate that female smokers have a higher likelihood of developing CHD compared to males. We observe that smokers with $cigsPerDay\ <\ 5$ are the biggest group, therefore, we observe that the latter is the group having the highest chance of getting a CHD, despite males and females having some numbers of getting CHD across all *cigsPerDay* interval cases.

Figure $7\ (b)$ shows the prediction counts for CHD over the range $\langle glucose,\ diaBP,\ sysBP \rangle$. The predictions reveal that patients with high levels of all three attributes ($glucose\ >\ 110$, $diaBP\ >\ 90$, and $sysBP\ >\ 180$) are at a significantly increased risk of developing CHD. While the risk decreases when *glucose* levels fall below 110; however, we observe that patients with high *diaBP* and *sysBP* values remain at risk, even if their *glucose* levels are low.

Figure 8 shows the prediction counts for CHD over the range $\langle age,\ education,\ totChol \rangle$. The predictions reveal that patients with lower education levels (degrees 1 and 2), as well as higher cholesterol levels ($totChol\ >\ 200$), tend to be connected with an increase in developing CHD, particularly among the older patients ($age\ >\ 50$). Contrarily, for patients under 40, the counts for CHD are relatively low across all *education* and *cholesterol* levels. It should be noticed here that the latter situation is very close to the real and effective today's status about health guidelines for large communities, at various scales such as the province scale and the regional scale.



(*a*) (*b*)

Figure 7: Prediction Counts for $\langle cigsPerDay,\ heartRate,\ gender \rangle\ (a)$ and $\langle glucose,\ sysBP,\ diaBP \rangle\ (b)$ (*Deep Learning* Network Model)

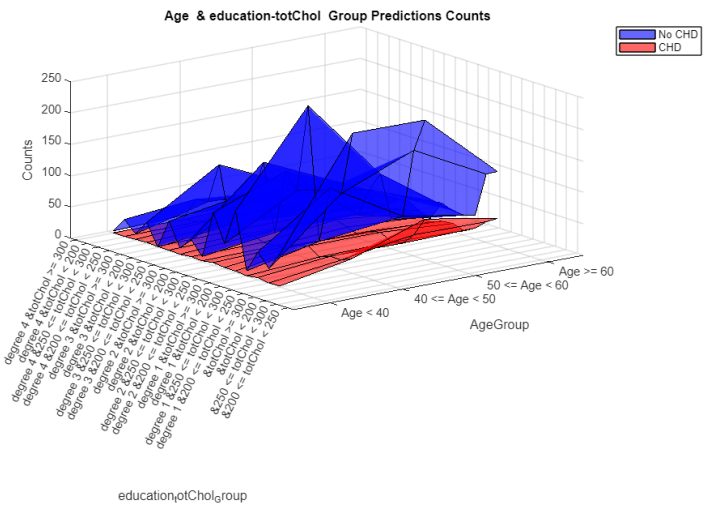


Figure 8: Prediction Counts for $\langle age, education, totChol \rangle$ (*Deep Learning Network* Model)

8. Conclusions and Future Work

This paper has explored the convergence of two innovative and very relevant scientific areas: digital twins and big data technologies. In particular, we focused on the issue of supporting multidimensional risk analysis and prediction over digital twin data, by nicely applying an emerging state-of-the-art analytics paradigm, the so-called multidimensional big data analytics.

The result is represented by a comprehensive framework, for which we provided anatomy and main functionalities. In addition to this, we conducted an extensive experimental evaluation and analysis of the performance of the proposed framework in both the different analytics and prediction phases, over state-of-the-art datasets populating the target healthcare digital twin scenario.

Future work is mainly oriented to further extend our contributions with specific aspects of the emerging big data trend, such as *performance optimization* and *outlier management*, and, also, studying possible integration with emerging flexible security paradigms (e.g., [46,47]).

Acknowledgements

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. The authors are grateful to Ilies Merzouk for his contribution to the experimental analysis of this research.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] F. Tao, B. Xiao, Q. Qi, J. Cheng, P. Ji, "Digital Twin Modeling". *Journal of Manufacturing Systems 64*, pp. 372-389, 2022

[2] L. Li, B. Lei, C. Mao, "Digital Twin in Smart Manufacturing". *Journal of Industrial Information Integration 26*, art. 100289, 2022

[3] Q. Qi, F. Tao, T. Hu, N. Anwer, A. Liu, Y. Wei, A. Y. Nee, "Enabling Technologies and Tools for Digital Twin". *Journal of Manufacturing Systems 58*, pp. 3-21, 2021

[4] F. Tao, J. Cheng, Q. Qi, M. Zhang, H. Zhang, F. Sui, "Digital Twin-Driven Product Design, Manufacturing and Service with Big Data". *The International Journal of Advanced Manufacturing Technology 94*, pp. 3563-3576, 2018
[5] Q. Qi, F. Tao, "Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison". *IEEE Access 6*, pp. 3585-3593, 2018
[6] L. Nie, X. Wang, Q. Zhao, Z. Shang, L. Feng, G. Li, "Digital Twin for Transportation Big Data: A Reinforcement Learning-Based Network Traffic Prediction Approach". *IEEE Transactions on Intelligent Transportation Systems 25(1)*, pp. 896-906, 2023
[7] L. Nele, G. Mattera, E.W. Yap, M. Vozza, S. Vespoli, "Towards the Application of Machine Learning in Digital Twin Technology: A Multi-Scale Review". *Discover Applied Sciences 6*, art. 502, 2024.
[8] S.M. Zayed, G. Attiya, A. El-Sayed, A. Sayed, E.E. Hemdan, "An Efficient Fault Diagnosis Framework for Digital Twins Using Optimized Machine Learning Models in Smart Industrial Control Systems". *International Journal of Computational Intelligence Systems 16(1)*, art. 69, 2023
[9] P. Calvo-Bascones, A. Voisin, P. Do, M.A. Sanz-Bobi, "A Collaborative Network of Digital Twins for Anomaly Detection Applications of Complex Systems. Snitch Digital Twin Concept". *Computers in Industry 144*, art. 103767, 2023
[10] W.Y. Wang, J. Yin, Z. Chai, X. Chen, W. Zhao, J. Lu, F. Sun, Q. Jia, X. Gao, B. Tang, X. Hui, H. Song, F. Xue, Z. K. Liu, J. Li, "Big Data-Assisted Digital Twins for the Smart Design and Manufacturing of Advanced Materials: From Atoms to Products". *Journal of Materials Informatics 2(1)*, pp. 1-27, 2022
[11] S. M. Zayed, G.M. Attiya, A. El Sayed. E.E. Hemdan, "A Review Study on Digital Twins with Artificial Intelligence and Internet of Things: Concepts, Opportunities, Challenges, Tools and Future Scope". *Multimedia Tools and Applications 82(30)*, pp. 47081-47107, 2023
[12] Y. Liang, Z. Yin, L. Nie, Y. Ba, "Shared Steering Control with Predictive Risk Field Enabled by Digital Twin". *IEEE Transactions on Intelligent Vehicles 8(5)*, pp. 3256-3269, 2023
[13] L. Lin, H. Bao, N. Dinh, "Uncertainty Quantification and Software Risk Analysis for Digital Twins in the Nearly Autonomous Management and Control Systems: A Review". *CoRR abs/2103.03680* 2021
[14] RiskDataControl.com, *The Role of Data Analytics in Risk Management*. Available at https://www.riskdatacontrol.com/role-data-analytics-risk-management/
[15] M. El Khatib, A. Ankit, I. Al Ameeri, H. Al Zaabi, R. Al Marqab, H.M. Alzoubi, M. Alshurideh, "The Role and Impact of Big Data in Organizational Risk Management". In: *The Effect of Information Technology on Business and Marketing Intelligence Systems. Studies in Computational Intelligence 1056*, Springer, pp. 2139-2153, 2023
[16] R. Gul, M. A. S. Al-Faryan, "From Insights to Impact: Leveraging Data Analytics for Data-Driven Decision-Making and Productivity in Banking Sector". *Humanities and Social Sciences Communications 10(1)*, pp. 1-8, 2023
[17] G. Dicuonzo, F. Donofrio, G. Galeone, "Big Data and Artificial Intelligence to Support Risk Management: A Systematic Literature Review". In: *Chiucchi, M.S., Lombardi, R., Mancini, D. (eds) Intellectual Capital, Smart Technologies and Digitalization. SIDREA Series in Accounting and Business Administration*, Springer, pp. 119-130, 2021
[18] A. Cuzzocrea, "OLAPing Big Social Data: Multidimensional Big Data Analytics over Big Social Data Repositories". In*: 4th International Conference on Cloud and Big Data Computing,* pp. 15-19, 2020
[19] A. Cuzzocrea, "Multidimensional Big Data Analytics over Big Web Knowledge Bases: Models, Issues, Research Trends, and a Reference Architecture". In: *8th IEEE International Conference on Multimedia Big Data*, pp. 1-6, 2022
[20] A. Cuzzocrea, "A Reference Architecture for Supporting Multidimensional Big Data Analytics over Big Web Knowledge Bases: Definitions, Implementation, Case Studies". *International Journal of Semantic Computing 17(4)*, pp. 545-568, 2023
[21] Z. Lv, D. Chen, H. Feng, A.K. Singh, W. Wei, H. Lv, "Computational Intelligence in Security of Digital Twins Big Graphic Data in Cyber-physical Systems of Smart Cities". *ACM Transactions on Management Information Systems 13(4)*, art. 39, 2022

[22] G. Coorey, G.A. Figtree, D.F. Fletcher, V.J. Snelson, S.T. Vernon, D.S. Winlaw, S.M. Grieve, A.L. McEwan, J.Y.H. Yang, P. Qian, K. O'Brien, J. Orchard, J. Kim, S. Patel, J. Redfern, "The Health Digital Twin to Tackle Cardiovascular Disease - A Review of an Emerging Interdisciplinary Field". *NPJ Digital Medicine 5*, art. 126, 2022
[23] S.A. Shinde, P.R. Rajeswari, "Intelligent Health Risk Prediction Systems using Machine Learning: A Review". *International Journal of Engineering & Technology 7(3)*, pp. 1019-1023, 2018
[24] G. Kalra, R. Boadh, Y.K. Rajoria, P. Rajendra, "Prediction and Analysis of Health Risk by using Fuzzy Control System". *International Journal of Health Sciences 6(2)*, pp. 7511-7524, 2022
[25] Y. Cheng, F. Wang, P. Zhang, J. Hu, "Risk Prediction with Electronic Health Records: A Deep Learning Approach". In: *SIAM International Conference on Data Mining*, pp. 432-440, 2016
[26] A. De Benedictis, N. Mazzocca, A. Somma, C. Strigaro, "Digital Twins in Healthcare: An Architectural Proposal and Its Application in a Social Distancing Case Study". *IEEE Journal of Biomedical and Health Informatics 27(10)*, pp. 5143-5154, 2023
[27] C. Meijer, H.W. Uh, S. El Bouhaddani, "Digital Twins in Healthcare: Methodological Challenges and Opportunities". *Journal of Personalized Medicine 13(10)*, art. 1522, 2023
[28] E. Zio, L. Miqueles, "Digital Twins in Safety Analysis, Risk Assessment and Emergency Management". *Reliability Engineering & System Safety 246*, art. 110040, 2024
[29] F. Tao, H. Zhang, A. Liu, A.Y.C. Nee, "Digital Twin in Industry: State-of-the-Art". *IEEE Transactions on Industrial Informatics 15(4)*, pp. 2405-2415, 2019
[30] H. Gong, S. Cheng, Z. Chen, Q. Li, "Data-Enabled Physics-Informed Machine Learning for Reduced-Order Modeling Digital Twin: Application to Nuclear Reactor Physics". *Nuclear Science and Engineering 196(6)*, pp. 668-693, 2022
[31] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, H. Pirahesh, "Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab, and Sub Totals". *Data Mining and Knowledge Discovery 1(1)*, pp. 29-53, 1997
[32] C. Dobre, F. Xhafa, "Parallel Programming Paradigms and Frameworks in Big Data Era". *International Journal of Parallel Programming 42(5)*, pp. 710-738, 2014
[33] A. Cuzzocrea, "Improving Range-Sum Query Evaluation on Data Cubes via Polynomial Approximation". *Data and Knowledge Engineering 56(2)*, pp. 85-121, 2006
[34] A. Cuzzocrea, R. Moussa, G. Xu, "OLAP*: Effectively and Efficiently Supporting Parallel OLAP over Big Data". In: *3rd International Conference on Model and Data Engineering*, pp. 38-49, 2013
[35] Y. Maïzi, A. Arcand, Y. Bendavid, "Digital Twin in Healthcare: Classification and Typology of Models Based on Hierarchy, Application, and Maturity". *Internet Things 28*, art. 101379, 2024
[36] S. Chaudhuri, U. Dayal, "An Overview of Data Warehousing and OLAP Technology". *SIGMOD Record 26(1)*, pp. 65-74, 1997
[37] A. Cuzzocrea, F. Furfaro, G.M. Mazzeo, D. Saccà, "A Grid Framework for Approximate Aggregate Query Answering on Summarized Sensor Network Readings". In: *2004 On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops*, pp. 144-153, 2004
[38] B. Yu, A. Cuzzocrea, D.H. Jeong, S. Maydebura, "On Managing Very Large Sensor-Network Data Using Bigtable". In: *12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pp. 918-922, 2012
[39] X.C. Yun, G. Wu, G. Zhang, K. Li, S. Wang, "FastRAQ: A Fast Approach to Range-Aggregate Queries in Big Data Environments". *IEEE Transactions on Cloud Computing 3(2)*, pp. 206-218, 2015
[40] A. Bhardwaj, *Framingham Heart Study Dataset*. Available at https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset
[41] National Institute of Diabetes and Digestive and Kidney Diseases Research, *Diabetes Healthcare: Comprehensive Dataset-AI*. Available at https://www.kaggle.com/datasets/deependraverma13/diabetes-healthcare-comprehensive-dataset
[42] M.A. Bhat, *Lung Cancer*. Available at https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
[43] D. West, "Neural Network Credit Scoring Models". *Computers & Operations Research 27(11-12)*, pp. 1131-1152, 2000
[44] Aggarwal, C.C., 2018. "*Neural Networks and Deep Learning*". Springer.

[45] S. Bayraci, O. Susuz, “A Deep Neural Network (DNN) based Classification Model in Application to Loan Default Prediction”. *Theoretical and Applied Economics 4(621)*, pp. 75-84, 2019

[46] M.J.H. Faruk, H. Shahriar, M. Valero, F.L. Barsha, S. Sobhan, M.A. Khan, M.E. Whitman, A. Cuzzocrea, D.C. Lo, A. Rahman, F. Wu, “Malware Detection and Prevention using Artificial Intelligence Techniques”. In: *9th IEEE International Conference on Big Data*, pp. 5369-5377, 2021

[47] M. Masum, H. Shahriar, H. Haddad, M.J.H. Faruk, M. Valero, M.A. Khan, M.A. Rahman, M.I. Adnan, A. Cuzzocrea, F. Wu, “Bayesian Hyperparameter Optimization for Deep Neural Network-Based Network Intrusion Detection”. In: *9th IEEE International Conference on Big Data*, pp. 5413-5419, 2021