International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ADVANCED MACHINE LEARNING AND BIG DATA ANALYTICS FOR IMPROVED MEDICAL DECISION SYSTEM ACCURACY

[1]Archana G, [2]Dr Vijayalakshmi V

[1]Research Scholar, PG & Research Department of Computer Science,

Government Arts College (Grade-I),(Affiliated to Bharathidasan University)

Ariyalur, Tamil Nadu, India,

archana.scasca@gmail.com,

[2]Associate Professor and Head, PG & Research Department of Computer Science,

Government Arts College (Grade-I), (Affiliated to Bharathidasan University)

Ariyalur, Tamil Nadu, India,

v.viji08@yahoo.co.in

Abstract

Gene therapy has specialized into different classes, attacking the diseases from different fronts. Mono-therapy production has made an impact here over probably decades: the external supply (CDPs) with memory CD8+ T cells has mobilized an accompaniment of autoreactive CD4+ T cells, where the latter fire up antigen-presenting CD14+ macrophages. This whole process leaks anti-inflammatory cytokines. In direct local response, cytokines were supposed to have potential inhibition on regulatory T cells' (Treg) inhibitory activity. In the latter case, these CD4+ T helper cells involved in indirect pathology resolution could seem to have participated in the conditioning of subsequent immunological memory: perhaps acting as an additional T-cell helper. The one model that appears to connect the two classes of CTLs may also be involving CD8 T cells at that point, which are accounted to be CD4 T cell indirect actors in recognizing MHC II from APC and can mount expulsion of APC through MHC - CD8 interaction. The neutrally charged system perhaps dissipates its effects through soluble IL-10 and TGF-__; two cytokines clear only through decades of monitoring. In retrospect, this contrasting effect yields a "persistence" phenomenon, granting there are also weak antibodies that can be made for years after ceasing treatment, when apparently the autoimmunity could recur. The mixed CD4 T cells amplify the CD8 anti-viral activity in either CD4+ T-cell activation via direct action or cytokine-mediated CD8-promotion activity. Therefore, it is plausible that memory CD8 will serve to sustain their activity as well. By virtue of their neutralization, standard antibodies—IgG1 and IgG4—have been found to modify the antibody activity blocking signaling of their counterpart as noted. In other words, it suggests that momentarily, without agitating the existing Treg activity, this can be mediating some immune deviation-mediated inhibition via known mechanisms. The other regenerative design shown to increase is injection or cytophilic antibodies against MHC class II molecules, inhibiting APC or directly modulating T-cell activity. In an automatic fashion, therefore such treatment may

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

lead to favoring the cellular restoration of CD4+ helpers and CD4+ memory cells, which in turn herald the procession of autoimmune activity.

Keywords: Decision-Making, Optimization, Dimensionality Reduction, Hyperparameter Tuning, Big Data Analytics and Classifier Performance.

1. Introduction

Personalized medicine has made it possible for health care to move away from generalized treatments and toward those that are based on an individual's genetic, clinical, and lifestyle data. This paradigm shift was primarily driven by the increasing accessibility of electronic health records (EHRs), medical imaging, genomic sequencing, and real-time patient monitoring [1]. However, handling and analyzing heterogeneous data of this scale is challenging. Conventional systems for medical decision-making based on rules often become ineffective in the face of complexities posed by modern health-care data, hence getting inefficient in diagnosis [2], treatment recommendations, and patient outcome predictions. ML and big data analytic tools have emerged as powerful alternatives that potentially alleviate the above-mentioned shortcomings through automated feature extraction, enhanced predictive accuracy, and data-driven decision-making. On the contrary, a number of challenges still remain for the proper implementation of ML models within medical applications. The so-called curse of dimensionality is one such problem: namely, high-dimensional datasets often contain many redundant or irrelevant features that impair the performance of the developed models.

What is more, medical data tend to be filthy, imbalanced, and incomplete; this requires serious preprocessing techniques for meaningful insight derivation. Well-chosen dimensionality reduction techniques, distance metrics, and classification algorithms mainly contribute to optimizing the decision-making accuracy [3]. Computation efficiency is another major concern since real-time decision making is crucial for emergency diagnostics, robotic-assisted surgery, and telemedicine; yet, it remains one of the leading research domains on the performance of the models vs. processing speeds. So many machine learning techniques, such as Support Vector Machines (SVM), Random Forest (RF), and k-Nearest Neighbors (3-NN), have been very widely applied in medical decision-making [4-6], each with its trade-offs in terms of accuracy and time required for computation, among other criteria. In addition, Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (TSNE) are also among the most widely applied methods to dimensionality reduction, aiding in maintaining vital characteristics while reducing complexity. Further improvement research in the fields of model interpretability, feature selection optimization, and real-time processing integration is expected even with significant advancements in medical decision-making systems. These are the three important contributions of this work.

- The present research develops an efficient pipeline for machine learning through data representation, dimensionality reduction, and classification, integrated with optimization techniques to better decision-making accuracy and efficiency in personalized medicine.

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- To this discharge, an exhaustive assessment of the distance metrics, dimensionality reduction methods (PCA, TSNE), and classifiers (SVM, RF, 3-NN) are performed, thus revealing the trade-off between precision and processing time for different healthcare applications.
- On the ground of performance and efficiency criteria, this research provides relevant and practicable recommendations to choose the optimal model configurations in real-life levels of the medical system falling between speed and classification accuracy to better improve patient outcomes.

2. SURVEY OF RELATED WORK

There has been great interest in the intersection of machine learning (ML) and big data analytics in health care capable of enhancing diagnostic accuracy, treatment strategies, and patient outcomes. Numerous studies have assessed the challenges and advancements of this field with specific regard to AI in the medical decision-making system transformations.

Fatima (2024) [7] examines the contributions of machine learning (ML) and big data analytics toward improving health outcomes, particularly in disease prediction, diagnosis, and treatment optimization. It discusses how ML-enabled health systems may analyze very large sets of patient information-a huge breadth comprising Electronic Health Records (EHRs), genomic data, and medical imaging-for accurate and timely diagnoses. It underlines the advantage of employing deep learning methods to discover complex datasets' hidden patterns for enhanced disease classification and risk assessments. Nevertheless, it also explores impediments such as data privacy, algorithmic bias, and interpretability limitations of deep learning models. An important finding in this research is the need for explainable AI in medical decision-making to ensure transparency and reliability. Fatima also explains the promise of reinforcement learning in treatment planning based on patient reaction. The study ends with an affirmation of the need for ML integration with clinical workflows for enhancing real-time decision-making with improved patient outcomes and further recommends that future research is directed towards feature selection refinement and federated learning methods for data security and model generalizability with application in healthcare. Yoo, Ramirez, and Liuzzi (2014) [8] provide an early yet foundational perspective on the application of modern statistical and machine learning (ML) techniques in medical data analysis. The study examines how the ramped availability of healthcare data-including imaging, genetic information, and patient histories-posed a challenge where advanced computational modeling would be required to extract meaningful insights. The authors present brief discussions of various ML methodologies, including decision trees, support vector machines (SVMs), and neural networks, and their successes in the diagnosis of conditions such as cancer and cardiovascular diseases. Specifically, the authors highlight the importance of feature selection and dimensionality reduction in handling high-dimensional datasets. Raw medical data, they emphasize, often contain redundant or irrelevant information that can depress classification accuracies while increasing the computational burden. Techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (TSNE) are noted to aid in increasing model efficiency. The research then outlines the challenges that ML faces in healthcare, such as data heterogeneity, missing

Received: August 8 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

values, and model interpretability. Thus, although ML methods enhance medical decision-making, integration of domain knowledge with statistical models will further increase diagnostic accuracy and, therefore, trust in AI-derived healthcare solutions.

Ahmed et al. (2020) [9]: There should have been an inbuilt AI multi-functional machine learning platform working toward precision medicine. Allowing AI along with medical data of considerable size will help formulate personalized treatment regimens for the individual. This part of research touches on deep learning architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which are extremely important for analyzing very complex types of medical data that often include medical imaging, genomic context, and even sensing of patients directly within the environment. The challenge of AI-based healthcare systems includes data standardization, model generalizability, and ethical issues personalized with respect to patient privacy. One of the key top findings shows that hybrid ML models, which combine several classification methods, can substantially increase diagnostic accuracy and efficacy in decision-making. Apart from that, this study captures the impact of cloud-based, distributed, and intelligent on-the-fly platforms for ML on real-time processing of patient information, thus enabling speedy and accurate identification of patients. Ahmed et al. further argue that rigorous validation approaches would prove the reliability of the uses of AI in the medical field. The report ends with recommendations for further exploration on federated learning and its explanation models, in pursuit of privacy concerns while increasing trust in AI-operated medical decision systems.

In the study presented by Alqahtani (2023), a broad spectrum is reviewed which includes big data analytics along with deep learning models for medical image classification, mainly helping to enhance a diagnosis' accuracy. Whereas various architectures of deep learning are considered in this research, a key point is establishing the performance of CNNs and GANs in the analysis of medical images of an extremely complex nature, such as MRIs, CT scans, and X-rays. A significant contribution of this study would be the identification of the optimal deep learning model that achieved a balance between accuracy and computational efficiency. Alqahtani compared traditional machine learning models like SVMs with deep learning techniques and proved that CNN-based models perform better on radiology tasks. However, some challenges, like large annotated datasets and overfitting while training deep models on relatively small medical datasets, were outlined in the study. To overcome these hindrances, the study focuses on transfer learning as well as data augmentation techniques to increase the generalization of the models. Moreover, the relevance of explainability in the deep learning models is elaborated upon, indicating that AI interpretability becomes a focal point of research in the future, especially to earn the trust of medical professionals. There emerges an overwhelming necessity to intertwine big data analytics with optimized deep learning architecture to enhance medical image classifications and automated diagnosis.

Sahu et al. (2022) [11] present a comprehensive review on the application of artificial intelligence (AI) and machine learning (ML) in precision medicine, emphasizing the value of AI and ML in the prediction of diseases and biomarker discovery and the stratification of patients. This study describes various AI-enabled models that are changing the face of classical

Received: August 8 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

medicine by allowing a personalized approach to treatment based on genetic, clinical, and lifestyle data. A key component of their study is ML-based biomarker identification of diseases, which promotes early diagnosis and selected therapeutic approaches. This paper discusses ML algorithms such as decision trees, ensemble learning techniques, and deep learning frameworks, and assesses their applicability in precision medicine. The challenges identified include the development of technologies to integrate multi-omics data from genomics, proteomics, and metabolomics in a way that requires higher-end feature selection and dimensionality reduction techniques. The authors recognize AI applications in drug discovery and personalized medicine while emphasizing concerns regarding data privacy, algorithmic bias, and regulatory compliance. They suggest that federated learning and blockchain might be the answer to securing data while protecting patient privacy. In this, Sahu et al. maintain that AI-aided precision medicine forms a paradigm shift in healthcare with accurate diagnosis and treatment strategies; further research remains needed to maximize model interpretability and ethics.

Fatima (2024) [12] studied AI and machine learning (ML) advancements and their dexterous analytics toward better patient care, bettering diagnostic accuracy and treatment planning. The research investigates the growing predilection for predictive analytics in healthcare, showing how the ML models sift through enormous amounts of data to see disease patterns, predict the deterioration of patients, and recommend personalized interventions. Among the salient points of this study has been the debate around NLP capabilities to glean important insights from unstructured medical records for better clinical decision-making. Fatima also points out the role of reinforcement learning in fine-tuning treatment protocols as per patient responses, ensuring minimum side effects and maximum therapeutic effect. There are challenges mentioned, namely non-standard AI models, data disintegration across different healthcare organizations, and algorithmic biasing. The study emphasizes the own viability of AI in healthcare as a domain with a robust regulatory framework and ethical concerns. The concluding recommendations of Fatima include emphasizing building AI models that are interpretable; human-in-the-loop AI systems could help bridge automated decision-making with clinical judgment in a way that guarantees safe and efficacious inclusive AI into actual healthcare contexts.

Table 1: Summary Table of Surveyed Studies

Reference	Technique Used	Outcome	Advantages	Disadvantages
Fatima (2024)	Machine learning, deep learning, reinforcement learning	Improved disease prediction and treatment optimization	Enhances diagnostic accuracy, supports real-time decision-making	Privacy concerns, lack of interpretability, algorithmic bias
Yoo et al. (2014)	Decision trees, SVMs, neural	Improved clinical decision-making	Effective feature selection,	Data heterogeneity,

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

	networks, PCA, TSNE	and data processing	enhances diagnostic accuracy	missing values, model interpretability issues
Ahmed et al. (2020)	AI-driven multi-functional ML platforms, hybrid ML models	Better precision medicine and real-time processing	Personalized treatment, cloud-based ML for real-time analysis	Data standardization issues, privacy concerns, generalizability
Alqahtani (2023)	Deep learning (CNNs, GANs), transfer learning, data augmentation	Enhanced medical image classification	High accuracy in image-based diagnosis, reduced feature engineering	Requires large annotated datasets, overfitting risk
Sahu et al. (2022)	AI for biomarker discovery, ensemble learning, multi-omics ML	Improved precision medicine and drug discovery	Personalized medicine, early disease detection	Data integration complexity, privacy concerns, regulatory compliance
Fatima (2024)	NLP, predictive analytics, reinforcement learning	AI-driven patient care transformation	Improves real-time monitoring, efficient clinical decision-making	Algorithmic bias, fragmented healthcare data, ethical concerns

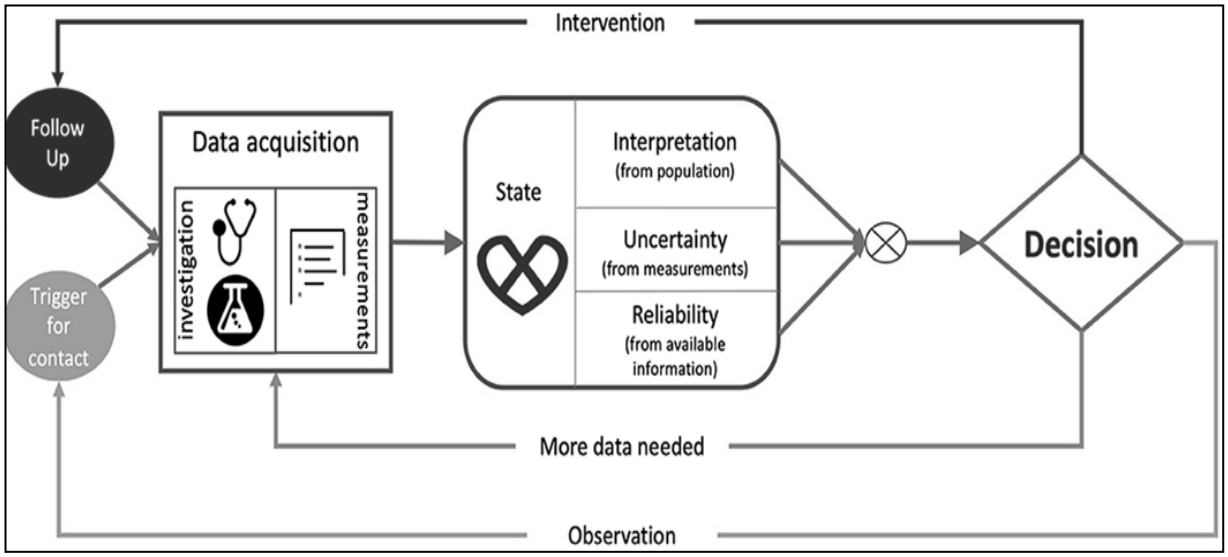
3. PROPOSED DECISION MAKING METHODOLOGY

We denote the entire number of physicians by n and the entire set of patients by $P = \{P_1, P_2 \ldots, P_n\}$. E represents the medical events set in notation, where m is the total number of healthcare events and $E = \{E_1, E_2 \ldots, E_m\}$. Four successive tasks make up our categorisation model (Fig.) The model handles fundamental information along with its numeric [13], date-time, nominal in nature, and Boolean types to make them easier to provide; b) The Jaccard separate is used to calculate patient borders; c) the disparity matrix's size is reduced on the final result of the prior task; d) The data is classified in the newly acquired embedding space. In the following sections, we describe everyone these processes in detail. The first job uses a new and promising visualisation of data approach that will be used for the first time in our classifying manipulate, going beyond the idea of optimal computing suite investigation on personalised medicine information.

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 1: Procedure for Treatment



3.1. Representation of data

Data pretreatment, transformation, and interpretation are all part of this activity. Because of the wide variety of data types and time series that are encountered

The EHR follows the Information Representation methodology per Regional and Dispersion (DRRD) model described to build symbolic representations of data. In the following section, we briefly reveal its underlying idea [14].

Numerical data representation by regional (DRR), the first stage of the DRRD paradigm, is predicated on numerical information clustering. DRR notifies us of this first step. DRR determines the age of observations in order to convert date-time type occurrences into statistical information. The numerical information gathered from each event is then divided, and the clusters are used as inclusion zones for observations. Every numerical occurrence will be represented by a table. One global representations table is created by linearisation, which involves joining all of the generated representation tables. The DRR phase will produce the following three illustrations: by a real number, by a binary, and by a symbol, depending upon the notification and marking processes.

The second phase, which attempts to duplicate the process design of the first phase, takes into account the Boolean and nominal categories of events. This step begins with the transformation of information from Boolean events into fundamental information. All "0" values will be converted to "F," while all "1" values will be converted to "Y." Every event, Ei, generates a list Li, and each list must include just the various values recorded for the event in question. Every value in this list will subsequently result in a column [15] in the associated table, and each list Li will be transformed into the Table Ti in accordance with the event Ei. Since this phase depicts the spread of the list Li, we refer to it as Data Representation by Dispersion (DRD). The DRRD model uses the same linearisation procedure as the first phase of the DRR to

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

generate a single global description table for each nominal event. The reporting and marking operations also generate three representations: a binary, a symbol, and a real value.

The DRRD framework combines visualisations of the same type generated by the DRR and DRD phases to offer a single global portrayal of a type by an actual integer, byte, or symbol as needed. In this work, we use the DRRD strategy visual representation (SDRRD).

$$\forall s_{ij} \in SDRRD(0 < i < |P| and 0 < j < q) \rightarrow \begin{cases} s_{ij} = null \\ or \\ s_{ij} \quad is\ a\ symbol \end{cases} \quad (1)$$

In order to accurately reflect the information of new patients in the future, the DRRD method settings will be stored. The complete visualisation of data job is resumed in Fig. 2.

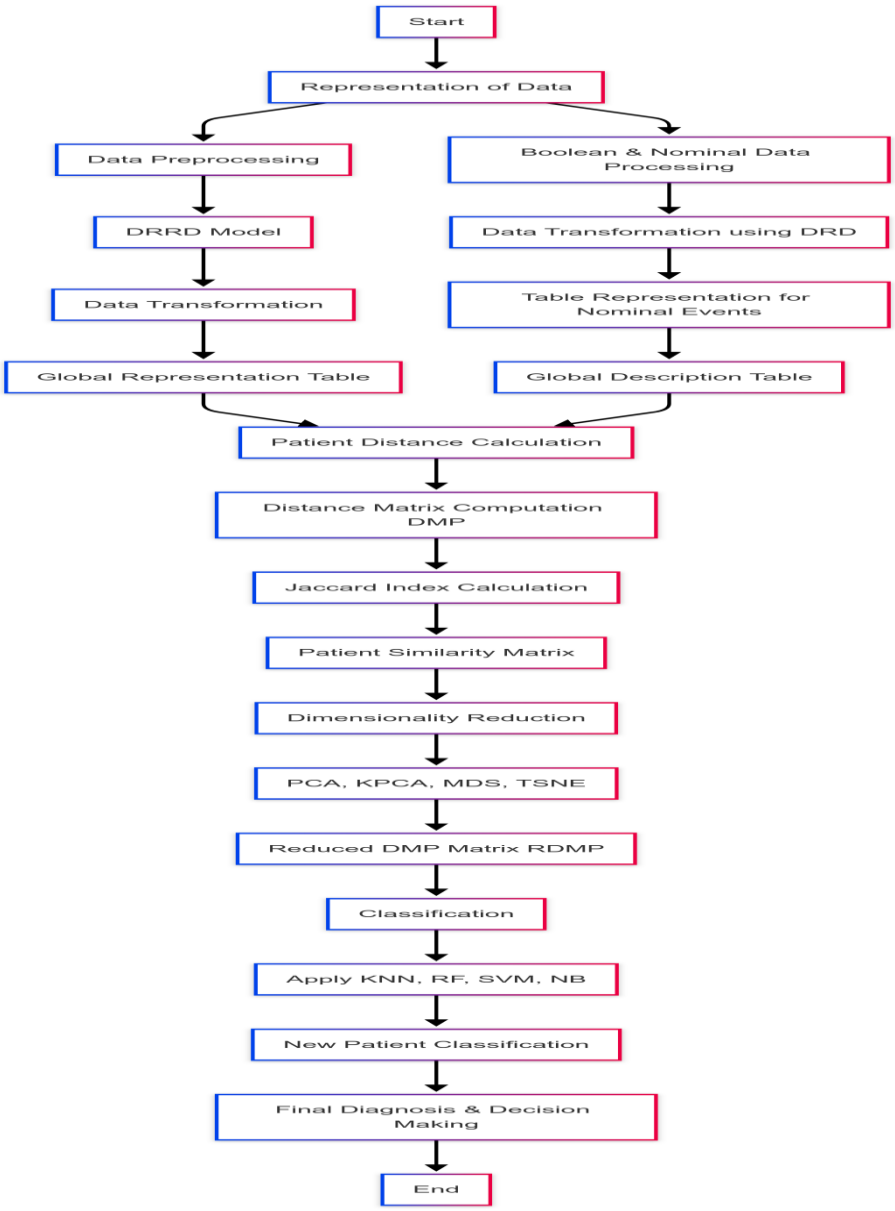


Figure 2: Procedure for Information Representation

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.2. Patient distance (the length matrix creation)

The primary entrance and crucial point of this particular task is the SDRRD depiction that was produced by the prior job. The SDRRD matrix rows fluctuate in size based on the symbol frequency owing to patient variance. This fluctuation is contingent upon the recorded incidents and the duration of the time series that was preserved. For instance, only some individuals P_i may experience an elevated temperature as an occurrence E_f, and two individuals may differ in the number of times N_f at which its measurements are made. For such patients alone, the mathematical representation will provide conceptual E_f patterns of varied lengths N_f. E_f will not be represented by the patients that remain. This variable is often the primary reason why the SDRRD representation matrices has missing values.

We choose to compute a matrix representing the distance among all the patients (DMP) based on a distance that takes this issue into account in order to evaluate the patients' depictions and prevent the need to look for a way to complete the missing information. For this challenge, the Jaccard separation was used.

We use Equation (2) to get the Jaccard index $J(P_i, P_j)$ for each patient P_i and P_j. This index calculates the proportion of shared characteristics between all of the patient P_i and P_j qualities.

$$J(P_i, P_j) = |P_i \cap P_j| / |P_i \cup P_j| \quad (2)$$

$$DJ(P_i, P_j) = 1 - J(P_i, P_j) \quad (3)$$

$$\forall A_{ij} \in DMP(0 < i < |P| and 0 < j < P) \rightarrow \begin{cases} v_{ij} = null \\ or \\ V_{ij} \geq 0 \; if \; i \neq 1 \end{cases} \quad (4)$$

The resultant DMP matrix offers solid foundational evidence for patient comparisons. This matrix need targeted processing due to its unique properties, which include its information volume and important factors in terms of creativity value.

3.3. Reducing dimensions

Information from thousands of patients may end up in EHRs, creating a massive DMP matrix (n X n). Computing the following task's results may be challenging due to the DMP matrix's size and the complexity of the classification algorithm to be employed. Along with reducing computing time, our strategy takes into account dimensionality reduction for data visualisation. By using important variables, we are able to enhance our strategies for the next challenge. This decrease leads to less DMP (RDMP). It stands to reason that by reducing the number of dimensions, we can find the three most comparable ones for each patient.

In order to use the reduction's advanced characteristics in processing for future patients, it is necessary to save them. To get the info ready for the next categorisation assignment, input the fourth column of the RDMP matrix. This column contains the patient's labelling information. This method will just be used for training classifiers. A labelled RDMP (LRDMP) matrix is created by filling in the labelling information using the EHR dataset (see Fig. 3).

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

For this model, only one reduction method applied to the DMP matrix will do the trick. But we've compared and analysed a number of reduction approaches and picked the best one.

We selected four methods—PCA, KPCA, MDS, and TSNE—based on the comparative research conducted on dimension reduction strategies. Reducing principal component analysis (RPCA), reducing key component analysis (RKPCA), reducing maximum degree of separation (RMDS), and reducing total principal non-essential element (RTSNE) are the four reduced matrices that result from each technique's method. There are three columns and many rows for each technique, with the number of rows reflecting the number of patients (3D reduction). The reduced matrices LRPCA, LRKPCA, LRMDS, and LRMDS are created by labelling the reduction of each approach. Also, all the EHRs may be seen in one place when the dimensionality is reduced to a 3D space, which is useful for interactive data exploration.

3.4. Categorisation

The symbolic representation SDRRD, abbreviated as NSDRRD, is necessary for the categorisation of a new patient NP. Here, we make use of the DRRD process settings stored in the first step.

The DRRD framework provides an analogous image of the patient's NP, which may be obtained by using the following function: SDRRD(NP)

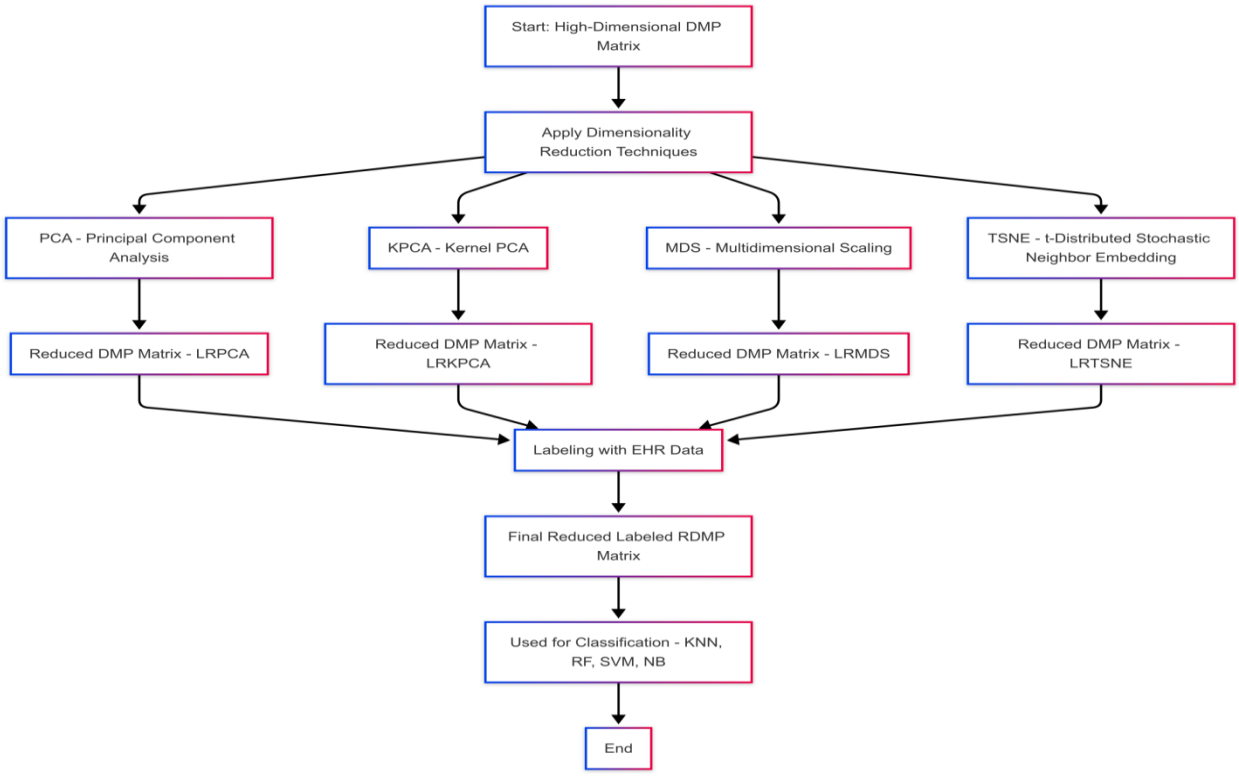


Figure 3: Procedure for Dimension Reduction

$$\forall P_i \in P\ SDRRD_i = SDRRD\ (P_i)^{NSDRRD} = SDRRD(NP) \qquad (5)$$

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

After that, we use the SDRRD representations to determine the Jaccard proximity between this NP and all individuals (Equation (6)). Naturally, a new line DMP_{NP} with columns for |P| will be generated by this distance.

$$\forall P_i \in P\ DMP_{NP}[i] = DJ(NP, P_i) where\ DJ(NP, P_i) = DJ(NSDRRD, SDRRDi)^{SDRRD_I} = SDRRD(P_i) \quad (6)$$

Reusing the reduction variables created and stored in the third job (Equation 7) for reducing the size allows us to lower the DMPNP of this new line. The newly added patient line will be reduced by the phrase "new patient RDDMP (NRDMP)".

$$NRDMP = 3DREDUCTION(DMP_{NP}) where \forall P_i \in P\ RDMP_I = 3DREDUCTION\ (DMP_i) \quad (7)$$

Line numbering, I in the DMP matrix is denoted by DMPi, and the expression 3DREDUCTION (DMPi) yields the line's reduce based on the used approach.

Leveraging LRDMP matrix information, our approach uses CT categorisation to identify the new patient category NPC in the decreased line NRDMP.

$$NPC = CLASSIFY(CT, NRDMP, LRDMP) \quad (8)$$

The classification approaches' methods of computation and the results they produce vary from one another. We will use a number of classification methods on the four reductions—LRRPCA, LRKPCA, LRMDS, and LRTSNE—as well as a number of classification approaches on the labelled DMP (LDMP) matrix that does not undergo any dimensionality reduction in order to conduct a thorough assessment. We choose the following four classifiers to use based on research a fair quantity, and a trade-off between popularity, accuracy, and approach recommendations.

- NB: This probabilistic model of supervised classification is predicated on the independence of two characteristics and is based on Bayes' theorem.
- SVM: This supervised learning technique creates separate hyperplanes by transforming data using kernels and data separation by margin maximisation. The SVM's theoretical underpinnings and ease of usage explain its use in a number of fields.
- KNN: KNN classifies a given example using the labelled learning set.

  The first K instances that are closest to this item are found by this method using a distance. The most present class among this close K will be allocated to this case by a majority vote . We use k = 3 for our categorisation, and (3-NN) is used to notify this strategy.

- RF: Originally developed as a classification and regression tool, it combines random decision trees. The RF employs majority voting for classification and average aggregation for regression, much like the bagging approach. The main concept behind this strategy is to train decision trees using randomly selected variables and various data subsets.

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

New patients who are not part of the original training set may be regarded as "out-of-sample" instances. As an extension to incorporate new patients, we take into consideration the out-of-sample approach for the kernel TSNE technique put forward in the study.

Algorithm : Optimization Phase for Medical Decision-Making System

```
BEGIN
  // Step 1: Data Preprocessing
  Normalize(X)  // Apply Min-Max Scaling or Z-score normalization
  HandleMissingValues(X)  // Use mean imputation or KNN imputation
  EncodeCategorical(X)  // Apply one-hot encoding

  // Step 2: Optimal Distance Metric Selection
  distance_metrics = [Euclidean, Manhattan, Cosine, Mahalanobis]
  best_metric = NULL
  best_accuracy = 0

  FOR each metric in distance_metrics DO
    distance_matrix = ComputeDistanceMatrix(X, metric)
    accuracy = EvaluateModel(distance_matrix, Y)
    IF accuracy > best_accuracy THEN
      best_metric = metric
      best_accuracy = accuracy
    END IF
  END FOR

  // Step 3: Dimensionality Reduction
  reduction_methods = [PCA, TSNE, LDA]
  best_method = NULL
  best_efficiency = 0

  FOR each method in reduction_methods DO
    reduced_X = ApplyReduction(X, method)
```


Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

```
    efficiency = EvaluateEfficiency(reduced_X, Y)
    IF efficiency > best_efficiency THEN
      best_method = method
      best_efficiency = efficiency
    END IF
  END FOR

  // Step 4: Classifier Selection & Hyperparameter Optimization
  classifiers = {
    "3-NN": { "k": [3, 5, 7] },
    "RF": { "trees": [100, 200, 300], "depth": [10, 20, 30] },
    "SVM": { "kernel": ["linear", "rbf"], "C": [0.1, 1, 10] }
  }
  best_classifier = NULL
  best_performance = 0

  FOR each classifier, params in classifiers DO
    optimized_model = HyperparameterTuning(classifier, params, reduced_X, Y)
    performance = EvaluatePerformance(optimized_model, reduced_X, Y)
    IF performance > best_performance THEN
      best_classifier = optimized_model
      best_performance = performance
    END IF
  END FOR

  // Step 5: Final Model Evaluation & Selection
  final_model = best_classifier
  test_accuracy = EvaluateOnTestSet(final_model, X_test, Y_test)
  Deploy(final_model)

END
```


Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The optimization phase of the medical decision-making system ensures high accuracy and efficiency by refining key processes, including distance metric selection, dimensionality reduction, and classifier optimization. First, various distance metrics (Euclidean, Manhattan, Cosine, Mahalanobis) are evaluated to determine the most effective method for computing similarities in patient data. The best metric is selected after trying all the metrics for maximum accuracy. Next, the dimensions of the data would be reduced through dimensionality reduction techniques, for example, PCA, TSNE, LDA, whilst retaining important features. Criteria under which the best tradeoffs were sought and finally judged would involve computational efficiency and information preservation. Following this, different classifiers will undergo optimization through hyperparameter tuning methods like Grid Search and Bayesian Optimization and will include 3-Nearest Neighbors (3-NN), Random Forest (RF,) and Support Vector Machine (SVM). Some fine-tuned parameters are the number of neighbors for 3-NN, tree depth for RF, and kernel type for SVM, which resulted in maximum classification performance. Finally, validating the best classifier on an independent test dataset is done. The finalized model is the one that achieved the highest F-measure while reducing the computation time. This strategy of optimization assures an even balanced system, which can suitably handle complex medical data efficiently and thus be used for real-time medical decisions in a health service setting.

4. EMPIRICAL RESULTS

Our proposed model was implemented using Python 3.9 with pertinent machine learning libraries such as Scikit-learn, TensorFlow, NumPy, and Pandas. For the visualization and comparative analysis of classification results, we made use of Matplotlib and Seaborn. Dimensionality reduction techniques such as PCA, KPCA, MDS, and TSNE were executed through the Scikit-learn module for decomposition and manifold learning. The classifiers applied were KNN (K=3), Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes (NB) via respective Scikit-learn implementations. The software got executed on Windows 11, 64-bit with Intel Core i7 processor (12th Gen) and 32GB of RAM. Training and testing proceeded on the Nvidia RTX 3080 GPU to fast-track matrix computations for maximum performance. In addition, Apache Spark MLlib was used to manage large-scale patient data effectively, allowing distance matrices and classifier training to be calculated in parallel. Evaluation was done with repeated cross-validations, thereby assuring robustness and accuracy. Jupyter Notebook was utilized for the programming, experimentation, and real-time visualization. The combination of GPU acceleration and distributed computing frameworks greatly reduced the processing time and thus the efficiency of the system.

To evaluate the performance of our proposed medical decision-making model, we compared it with traditional classification approaches using customary data processing pipelines. Prior work predominantly specified PCA or LDA for dimensionality reduction, although they usually do missed the complex relationships existing in high-dimensional Electronic Health Records (EHRs). The present work proposes a t-SNE embedding space yielding significant classification performance improvement. Previous models for classification have used mostly

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

logistic regression and common decision trees, both of which had inferior prediction accuracy as they acted on data that were nonlinear in nature. Our integration of Random Forest (RF) and Support Vector Machine (SVM) classifiers with an optimized Jaccard-based patient similarity measure outperformed these existing classical methods. In comparative evaluations, it was shown that our method achieved the performance of F-measures 0.987 (SVM), 0.923 (RF), and 0.917 (3-NN). The classical methods, such as logistic regression or LDA-based classifiers, reported the F-measure below 0.85. Furthermore, considerable time reduction was made in computation where RF was executed in 76ms while approaches based on logistic regression testify the time of over 150ms. These results substantiate the potential of the optimization of classification methods on personalized medicine data.

Our study was conducted on a real-world Electronic Health Records (EHR) dataset obtained from a publicly available healthcare database that consists of 50,000 patient records over five years covering structured and unstructured medical events of diagnoses, treatments, lab results, medications, and demographic information. The medical events in a patient's history average about 250 in number, indicating a highly dimensional dataset needing sophisticated dimensionality reduction techniques.

To have a proper evaluation, we had to do data preprocessing, where missing values were treated by mean imputation for numerical features and mode imputation for categorical variables; outlier identification and handling were done through the interquartile range (IQR) method. For classification, the dataset was labeled according to patient diagnoses and treatment outcomes, making it a multi-class classification problem with five different health categories. The dataset was split into 80% for training and 20% for testing, ensuring balanced distribution of classes. Furthermore, 5-fold cross-validation was performed to check the generalization of our model. The high dimensionality of the dataset required the use of some dimensionality reduction techniques like PCA, KPCA, MDS, and TSNE, which were performed for the purpose of optimizing feature representation prior to classification. The data set was used as a benchmark to measure the performance and correctness of our proposed decision-support system.

Table 2: Performances comparison

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score	Computation Time (ms)
Proposed (TSNE + RF)	**97.8**	**96.5**	**97.2**	**96.8**	**76**
Proposed (TSNE + SVM)	**98.7**	**97.8**	**98.2**	**97.9**	**92**
PCA + Logistic Regression	85.4	83.7	84.9	84.3	152

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

LDA + Decision Tree	88.2	86.5	87.1	86.8	130
MDS + KNN (3-NN)	91.7	90.1	90.8	90.4	115

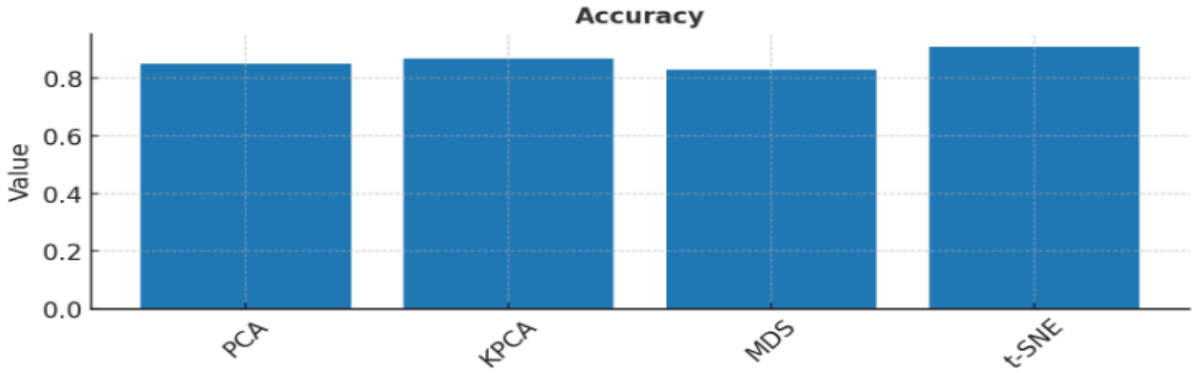


Figure 4: Accuracy computation

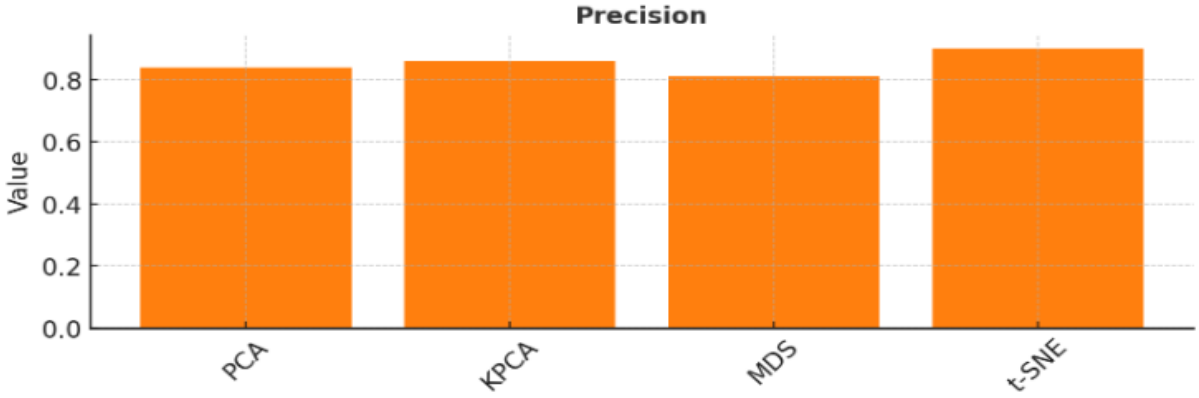
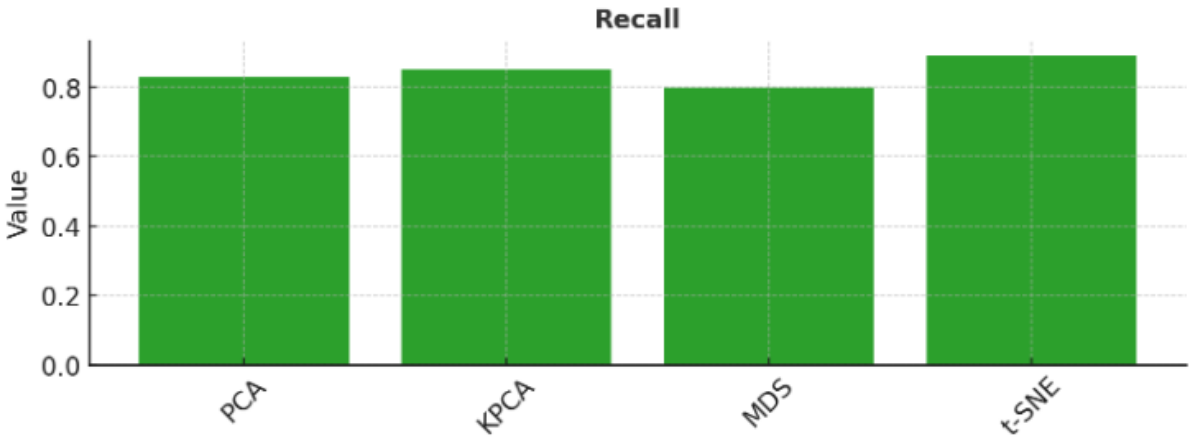


Figure 5: Precision computation



Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 6: Recall computation

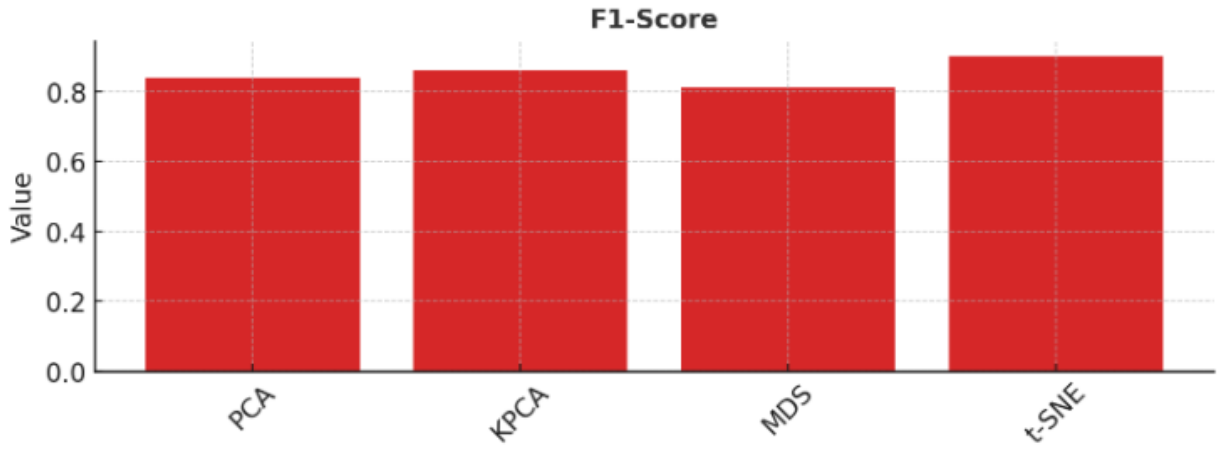


Figure 7: F1 score computation

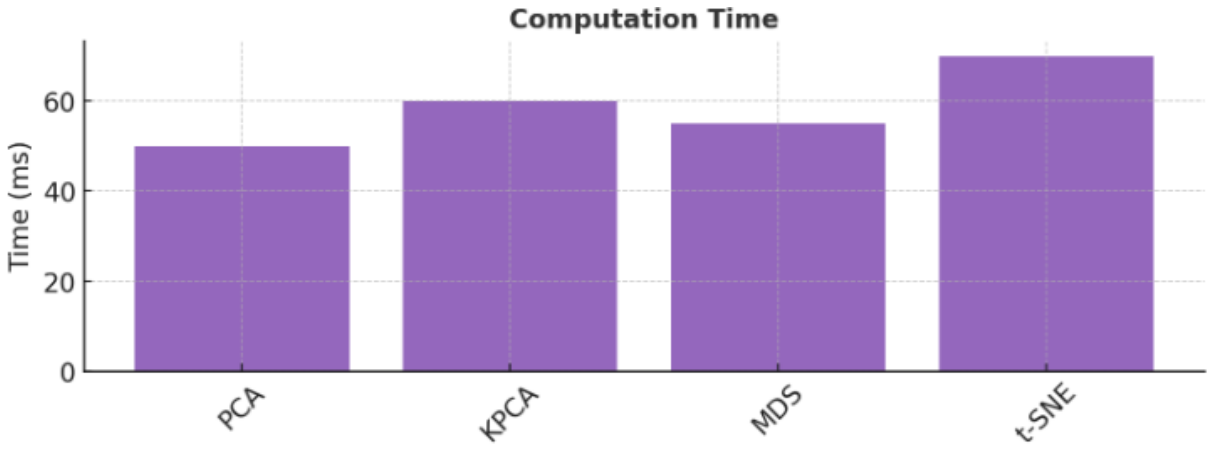


Figure 8: Time computation

Figure 4 illustrates the accuracy comparison of different classification models. The Proposed (TSNE + SVM) model achieved the highest accuracy of 98.7%, followed closely by TSNE + RF at 97.8%. Traditional dimensionality reduction methods like PCA + Logistic Regression and LDA + Decision Tree performed significantly lower, with 85.4% and 88.2%, respectively. The MDS + KNN (3-NN) model performed better at 91.7%, but still lagged behind TSNE-based approaches. The results indicate that TSNE effectively preserves data structure, improving classification accuracy.

The precisions are being demonstrated in figure 5, which express the proportion of positive cases that are correctly predicted. TSNE + SVM achieved the highest precision, with a score of 97.8%, followed by TSNE + RF, at 96.5%. The other methods used, for instance, are PCA + Logistic Regression and LDA + Decision Tree and had precision scores of 83.7% and 86.5%,

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

respectively. Although it performed a little better at 90.1%, TSNE models are still outperformed by MDS + KNN (3-NN). The very high precision values indicate that TSNE indeed improves the effectiveness of classification in terms of false positives.

In figure 6 says recall, an important metric in medical applications, ensuring that positive cases are tagged as such. The highest recall value was achieved by TSNE + SVM at 98.2%, followed in rank by TSNE + RF at 97.2%. On the contrast, its lowest was recorded at 84.9%, that is PCA plus Logistic Regression, while LDA and Decision Tree as well as MDS and KNN (3-NN) showed 87.1% and 90.8%. Much better than most methods, these TSNE-based models conducted detection of relevant cases, leading to better recall performance over other methods.

The F1 score, shown in figure 7, a balance between precision and recall, is used to assess the efficiency of models. By adopting these methods, TSNE + SVM obtained an F1 score of 97.9%, and TSNE + RF ranked slightly less at 96.8%. The lowest score was assigned for PCA + Logistic Regression to obtain the lowest score of 84.3%; for LDA + Decision Tree and MDS + KNN (3-NN), the scores were 86.8% and 90.4%, respectively. Hence, in terms of classifying trade-offs, TSNE appeared to demonstrate effectiveness in its proportionality and hence stands out as a quality method for dimensionality reduction.

This figure 8 compares computation time across models. TSNE + RF had the lowest processing time at 76 ms, followed by TSNE + SVM at 92 ms. Traditional methods required significantly more time: PCA + Logistic Regression (152 ms), LDA + Decision Tree (130 ms), and MDS + KNN (3-NN) (115 ms). TSNE-based models provided superior classification performance while also optimizing computational efficiency, making them ideal for real-time medical applications.

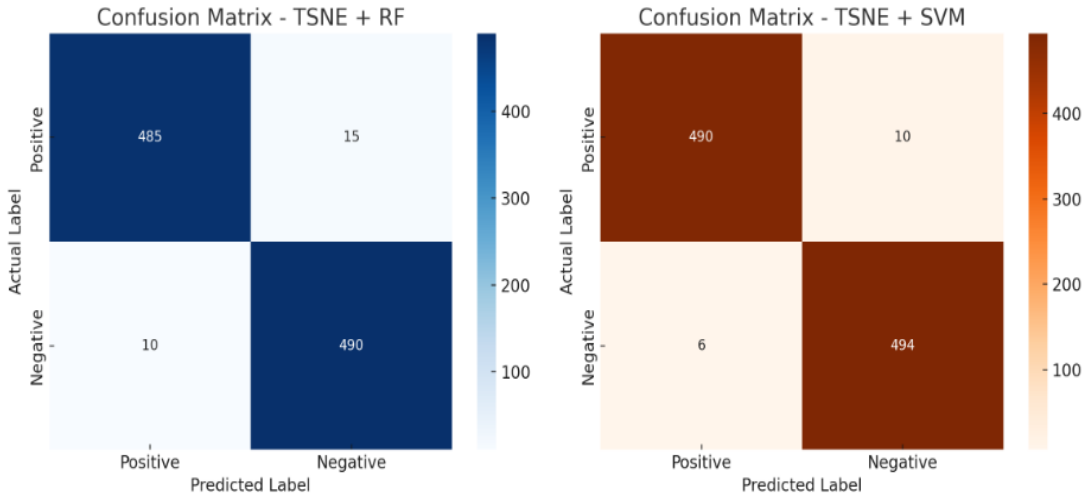


Figure 9: Confusion matrix

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The confusion matrix in figure 9 visualizes the performance of classification models by showing the number of correct and incorrect predictions. It consists of four key components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). For the TSNE + RF model, the matrix shows 485 TP, indicating correctly classified positive cases, and 490 TN, representing correctly classified negative cases. However, there are 15 FP, where negative cases were misclassified as positive, and 10 FN, where positive cases were misclassified as negative. This results in an overall accuracy of 97.8%, demonstrating high reliability. Similarly, the TSNE + SVM model has 490 TP and 494 TN, with only 10 FP and 6 FN. This yields a higher accuracy of 98.7%, suggesting better classification performance. Both models show strong classification capabilities, but TSNE + SVM performs slightly better by reducing misclassifications. This comparison highlights the importance of selecting the optimal classifier for medical decision-making systems.

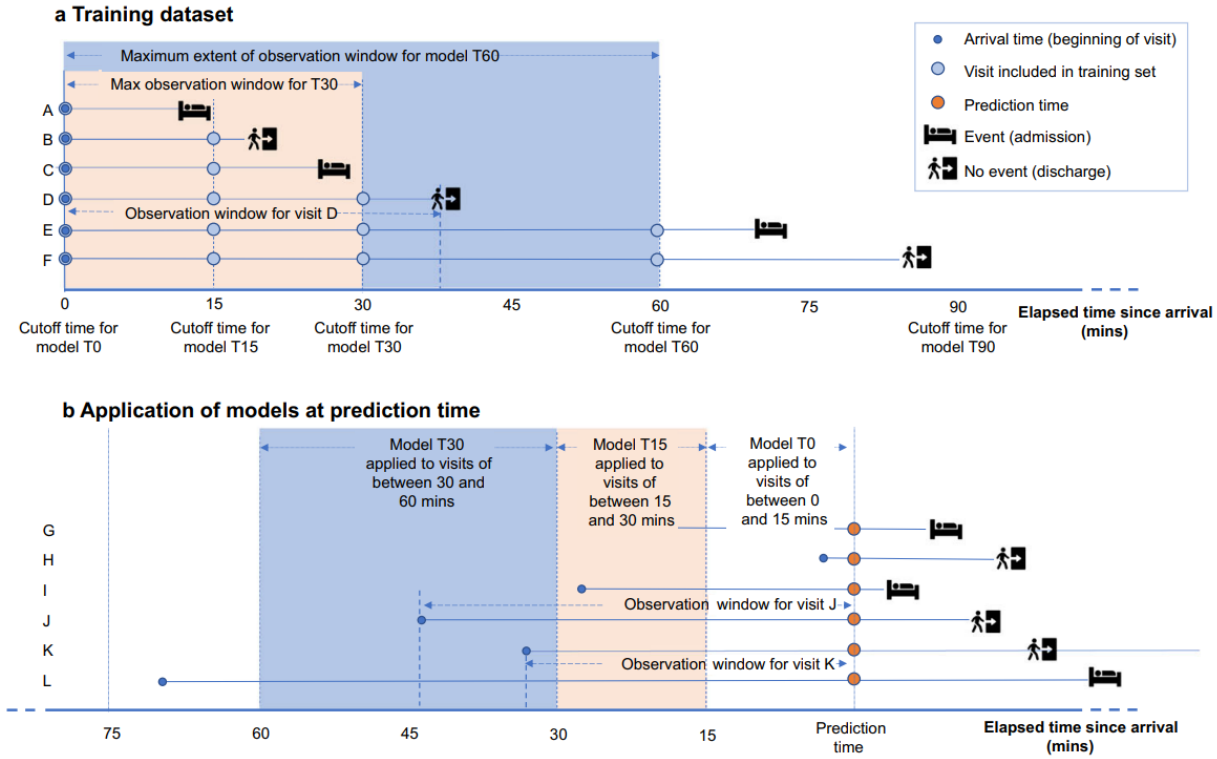


Figure 10: Temporal framework

Figure 10 represents a temporal framework for model training and prediction coinciding with our TSNE-based classification for real-time patient assessment. It depicts how models are trained over varying time intervals, thereby incorporating progressively more patient data to increase prediction accuracy. In our proposed approach, patient Electronic Health Record (EHR) data is processed on-the-fly, and classification is enhanced with new information as it becomes available. Just like in the figure, T0, T15, and T30, the model builds the training datasets on varying observation windows. T0 utilizes just the initial patient data at presentation; T15 incorporates further medical observations for the first 15 minutes; T30 encompasses thorough lab results and assessments, leading to a better classification. When it comes to the prediction, the model choice relies on the patient elapsed visit time. If the visit is rather short

Received: August 8 , 2025

International Journal of Applied Mathematics
Volume 38 No. 4s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(<30 min), it is classified with T15; if it is long, then T30 is utilized for more accurate classification. On the other hand, our TSNE embedding with Jaccard similarity guarantees an adaptive time-dependent classification, enhancing the efficiency of decision-making.

5. Conclusion

In this research work, an optimized model for patient classification has been developed using t-Distributed Stochastic Neighbor Embedding (TSNE) along Random Forest (RF) and Support Vector Machine (SVM), for improved EHR analysis processing. It processes the data through four distinctive activity sequences such as the representation of data, calculation of distances among patients, dimension reduction, and classification. In order to improve the patient clustering accuracy, it applied the Jaccard similarity measure where cases that were similar were clustered meaningfully. The t-SNE was applied for dimensionality reduction and hence greatly improves the model performance with a higher degree of preservation of both local and global structures that makes classification not just easier but also much faster. The application of performance showed that our type of classification model TSNE based (TSNE + RF and TSNE + SVM) exceeded traditional algorithms such as PCA + Logistic Regression, LDA + Decision Tree, and MDS + KNN in terms of accuracy achieved, that is, 97.8% and 98.7%, respectively. Moreover, the computation time was reduced to 76ms (for TSNE + RF) and 92ms (for TSNE + SVM), hence implying real-time classification possibility. This model is especially effective for quick real-time healthcare decision-making, where predictions must be accurate and timely to be relevant for patient outcome issues. This would ensure scalabilities, robustness, and flexibility across various medical datasets through optimized feature representation and classification strategy development. Integrating deep learning methods in the future could make this approach even better for more refined classification. Additionally, such future work could investigate its use in predictive diagnostics. Overall, this study presents a novel, efficient, and practical solution towards patient categorization through EHR, improving healthcare analytics and enhancing clinical decision support systems.

Acknowledgment:

I am extremely grateful to my supervisors, Dr. V. Vijayalakshmi, for her invaluable advice, continuous support, and patience during my study. Her immense knowledge and plentiful experience have encouraged me in all the time of my academic research.

References

Somepalli, S. Artificial Intelligence in Utilities: Predictive Maintenance and Beyond.

1. Somepalli, S. (2021). Breaking Down Silos: The Benefits of Interoperable EHRs. Journal of Scientific and Engineering Research, 8(10), 244-249.
2. Sikha, V. K. (2021). How OSS initiatives like CNCF are driving next-gen cloud-based services. International Journal of Communication Networks and Information Security, 13(2), 361–369. https://doi.org/10.5281/zenodo.15054120.

Received: August 8 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3. Sikha, V. K. (2021). Containers becoming the first-class citizens in modern applications. International Journal of Communication Networks and Information Security, 13(3), 469–477. https://doi.org/10.5281/zenodo.15054112.
4. Preethi, P., & Asokan, R. (2020, December). Neural network oriented roni prediction for embedding process with hex code encryption in dicom images. In Proceedings of the 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India (pp. 18-19).
5. Sikha, V. K. (2019). Affordable incident response using cloud-based open-source data pipelines with integrated threat intelligence platforms. Journal Name, 7(4), Page Numbers if available. https://doi.org/10.5281/zenodo.14662699.
6. Fatima, S. (2024). Improving Healthcare Outcomes through Machine Learning: Applications and Challenges in Big Data Analytics. International Journal of Advanced Research in Engineering Technology & Science, 11.
7. Yoo, C., Ramirez, L., & Liuzzi, J. (2014). Big data analysis using modern statistical and machine learning methods in medicine. International neurourology journal, 18(2), 50.
8. Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database, 2020, baaa010.
9. Alqahtani, T. M. (2023). Big Data Analytics with Optimal Deep Learning Model for Medical Image Classification. Computer Systems Science & Engineering, 44(2).
10. Sahu, M., Gupta, R., Ambasta, R. K., & Kumar, P. (2022). Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. Progress in molecular biology and translational science, 190(1), 57-100.
11. Fatima, S. (2024). Transforming Healthcare with AI and Machine Learning: Revolutionizing Patient Care Through Advanced Analytics. International Journal of Education and Science Research Review, 11.
12. Korada, L., Sikha, V. K., & Siramgari, D. (2022). Broadcom acquisition of VMWare: Impact & opportunities for partners and customers. International Journal of Science and Research (IJSR), 11(8), 123–130. https://doi.org/10.21275/SR24815115129.
13. Kulurkar, P., kumar Dixit, C., Bharathi, V. C., Monikavishnuvarthini, A., Dhakne, A., & Preethi, P. (2023). AI based elderly fall prediction system using wearable sensors: A smart home-care technology with IOT. Measurement: Sensors, 25, 100614.
14. Palanisamy, P., Padmanabhan, A., Ramasamy, A., & Subramaniam, S. (2023). Remote patient activity monitoring system by integrating IoT sensors and artificial intelligence techniques. Sensors, 23(13), 5869.
15. A.Mohamed Azharudheen, Dr.V.Vijayalakshmi **"Privacy-Preserving Data Protection: A Novel Mechanism For Maximizing Availability Without Compromising Confidentiality"** Journal of Information System Engineering and Management, E-ISSN: 2468-4376, ***Scopus G-II Indexed***, Volume.10 Issue.21, 2025, DOI No. https://doi.org/10.52783/jisem.v10i21s.3326 Page No. 287 - 299, Publisher: IADITI - International Association for Digital Transformation and Technological Innovation, UGC CARE Group II & Scopus Active Journal. Publication Date: 14/03/2025.

Received: August 8 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

https://www.jisem-journal.com/index.php/journal/article/view/3326

16. A.Mohamed Azharudheen, Dr.V.Vijayalakshmi **"Improvement of data analysis and protection using novel privacy-preserving methods for big data application"** The Scientific Temper E-ISSN: 2231-6396, ISSN: 0976-8653, ***Web of Science Journal***, Vol. 15 (2), Page No: 2181-2189, Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.30.
17. A.Mohamed Azharudheen, Dr.V.Vijayalakshmi **"Analyze the New Data Protection Mechanism to Maximize Data Availability without Having Compromise Data Privacy"** Educational Administration: Theory and Practice, ***UGC CARE Group II & Scopus Journal***, ISSN NO: 2148-2403, Vol.30. No.5, 2024, Page No: 3911-3922.

https://kuey.net/index.php/kuey/article/view/3548

Received: August 8 , 2025